(19) JAPANESE GOVERNMENT PATENT OFFICE

(11) Publication number: **0**

Generated Document.

PATENT ABSTRACTS OF JAPAN

(21) Application number: **05050398**

(22) Application date: **11.03.93**

(51) Intl. Cl.: **G06F 3/06** G06F 3/06

(30) Priority:

(43) Date of application publication: **22.09.94**

(84) Designated contracting states:

(71) Applicant: **HITACHI LTD**

(72) Inventor: **FUJII TETSUHIKO**
YAMAMOTO AKIRA
SATO TAKAO
YOSHIDA MINORU

(74) Representative:

(54) DISK ARRAY CONTROL METHOD

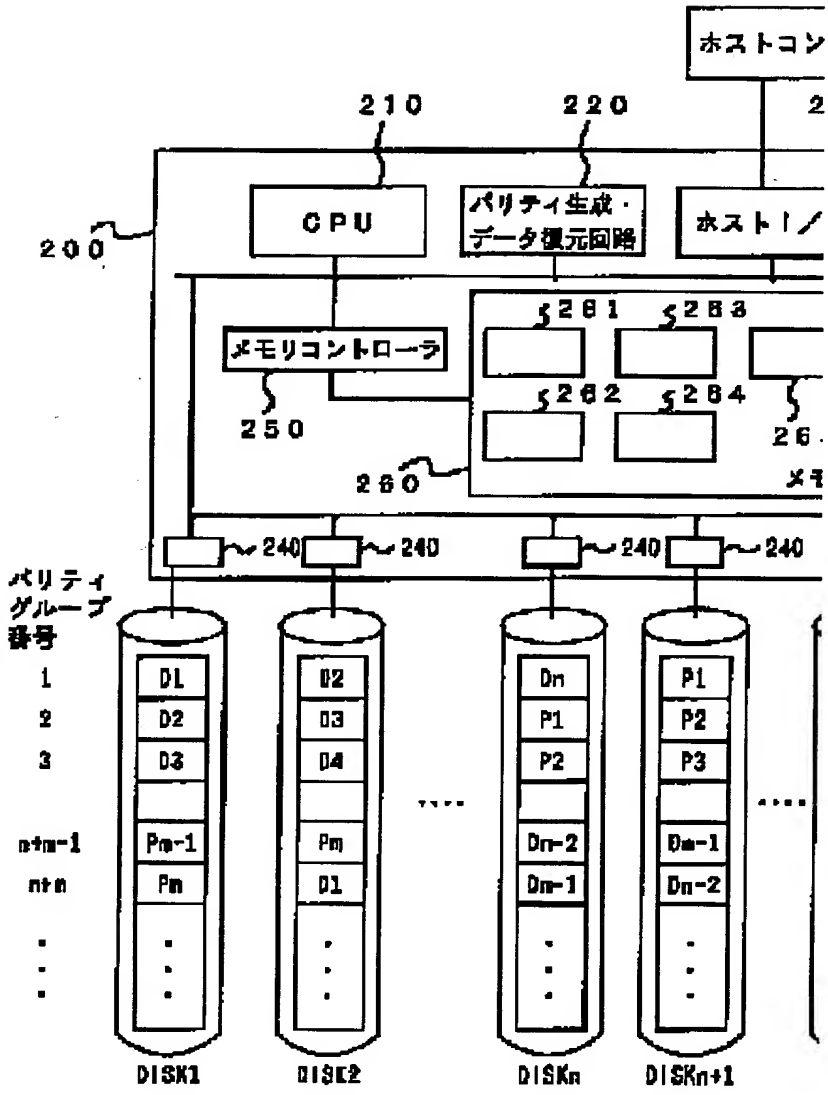
(57) Abstract:

PURPOSE: To prevent the increase of the frequency in access by using data and redundant data, which are stored in disk devices other than a disk device where a fault is detected, to restore data of this disk device and recording this data in the disk.

CONSTITUTION: If a parity group of a processing object is not restored yet, a classification (c) of data/redundant data stored in a faulty disk is obtained in this parity group. When the classification (c) indicates data, numbers of disks where data and redundant data belonging to the same parity group are stored are obtained from a parity constitution management table 264 for the purpose of restoring data, and they are read out from pertinent disks to restore data in the faulty disk. Restored data or generated redundant data is written over redundant data

selected for overwrite, and restoration state management information corresponding to the entry in the parity group of the processing object in a restoration state management table 263 is changed from the unrestored state to the already restored state to terminate the processing.

COPYRIGHT: (C)1994,JPO&Japio



(19)日本国特許庁（ＪＰ）　(12)公開特許公報(A)　(11)特許出願公開番号
特開平6－266508
(43)公開日　平成６年(1994)９月22日

(51)Int.Cl.[5]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　3/06	３０４　Ｂ	7165－5Ｂ		
	３０５　Ｆ	7165－5Ｂ		

審査請求　未請求　請求項の数11　ＯＬ　（全 27 頁）

(21)出願番号　特願平5－50398

(22)出願日　平成５年(1993)３月11日

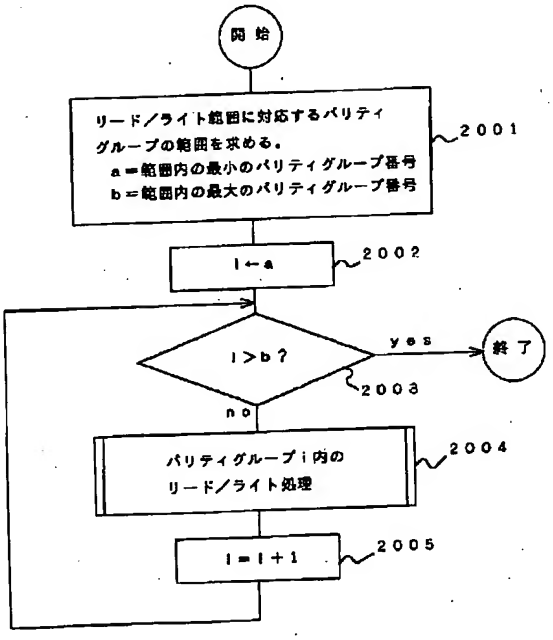
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　藤井　哲彦
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
(72)発明者　山本　彰
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
(72)発明者　佐藤　孝夫
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
(74)代理人　弁理士　小川　勝男

最終頁に続く

(54)【発明の名称】　ディスクアレイ制御方法

(57)【要約】
【目的】　冗長構成を採用したディスクアレイにおけるディスク故障時の性能低下を防止する。
【構成】　ディスクアレイを構成するディスクに故障が生じた際、故障ディスクに格納されていたデータを上書きする上書き用の冗長データを選択し、故障したディスク内のデータまたは冗長データを上書き用の冗長データの上に上書きし、ディスク構成を管理する管理テーブルを更新する。
【効果】　ディスク故障時のディスクアクセス回数を低減することができ、ディスク故障時のディスクアレイの性能低下を防止することができる。

図１９



1

【特許請求の範囲】

【請求項1】データを格納するための少なくとも1のディスク装置と冗長データを格納するための少なくとも1のディスク装置とを備えたディスクアレイシステムにおけるディスクアレイ制御方法において、前記データを格納するディスク装置および冗長データを格納するディスク装置の故障を検知し、前記故障が検知された場合、前記冗長データを格納するディスク装置から任意のディスク装置を選択し、前記故障が検知された以外のディスク装置に格納されたデータおよび冗長データを用い前記故障が検知されたディスク装置のデータを復元し、該復元されたデータを前記選択により選択されたディスクに記録することを特徴とするディスクアレイ制御方法。

【請求項2】前記復元するステップと、前記記録するステップは、ホストコンピュータからのディスクアレイへのリードおよびライトアクセスとは非同期なバックグラウンド処理として行なわれることを特徴とする請求項1記載のディスクアレイ制御方法。

【請求項3】前記復元するステップと、前記記録するステップは、ホストコンピュータからの故障ディスクへのリードおよび／またはライトアクセス時に同期して行なわれることを特徴とする請求項2記載のディスクアレイ制御方法。

【請求項4】請求項1記載のディスクアレイ制御方法において、前記ディスクアレイシステムは、故障ディスク装置を代替する少なくとも1台のホットスタンバイ用の予備ディスクを備え、前記故障を検知するステップにおいて故障が検知された際、前記呼びディスクが全て代替用として用いられている場合に前記選択するステップ、復元するステップ、および記録するステップを実行することを特徴とする請求項1記載のディスクアレイ制御方法。

【請求項5】前記各ステップを繰り返し実施することを特徴とする請求項1記載のディスクアレイ制御方法。

【請求項6】多重度n（≧1）のデータに対してm（≧1）個の冗長データを記録する、ディスクの冗長構成を採用したディスクアレイを制御するディスクアレイ制御方法であって、k（m≧k≧1）台のディスクが故障した時に、前記m個の冗長データの中からk個の冗長データを選択するステップと、故障したk台のディスク内に記録されていたk個のデータを前記ステップで選択したk個の冗長データの格納されていた領域に書き込むステップを有し、前記故障したk個のディスク内に記録されていたk個のデータを正常なディスク内の冗長データの格納されているk個の領域に書き込むことによりディスクの冗長構成を再構成することを特徴とするディスクアレイ制御方法。

【請求項7】前記ディスクの冗長構成の再構成処理が、ホストコンピュータから前記ディスクアレイへのリード／ライトアクセスとは非同期なバックグラウンド処理と

2

して行なわれることを特徴とする請求項6記載のディスクアレイ制御方法。

【請求項8】前記ディスクの冗長構成の再構成処理が、ホストコンピュータから故障ディスクへのリード／ライトアクセス時に、該リード／ライトアクセスと同期して行なわれることを特徴とする請求項6記載のディスクアレイ制御方法。

【請求項9】前記ディスクアレイは、さらにs（≧1）台のホットスタンバイの予備ディスクを有し、ディスク故障時に、全ての予備ディスクを使いきって、予備ディスクがなくなった状態で、更にディスクが故障したときに、前記ディスクの冗長構成の再構成処理を行うことを特徴とする請求項6記載のディスクアレイ制御方法。

【請求項10】前記ディスクアレイは、さらにs（≧1）台のホットスタンバイの予備ディスクを有し、k（m＋s≧k＞s）台のディスクが故障した時に、故障ディスク内のデータおよび冗長データの予備ディスクへの回復と前記ディスクの冗長構成の再構成とを行うことを特徴とする請求項6記載のディスクアレイ制御方法。

【請求項11】前記冗長データが2以上であって、前記ディスクの冗長構成の再構成を複数回繰り返すことを特徴とする請求項6記載のディスクアレイ制御方法。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、コンピュータシステムにおける外部記憶装置の制御方法にかかり、特に、複数のディスク装置により冗長構成をなすディスクアレイ装置の障害時の性能低下を防止を考慮したディスクアレイ制御方法に関する。

【0002】

【従来の技術】近年、コンピュータシステムにおけるI／Oネックを解決する為の、一つのアプローチとして、冗長構成のディスクアレイが注目され、技術開発が活発に行われている。冗長構成のディスクアレイはRAID（Redundant Array of Inexpensive Disks）と呼ばれ、その構成方式として、RAID3、RAID4、RAID5が知られている。RAIDに関しては、例えば、デー・パターソン、ジー・ギブソン、アール・カッツ等による、「ア ケース フォー リダンダントアレイズ オブ インイクスペンシブ ディスクス（レイド）、エーシーエムシグモッド カンファレンス プロシーディングズ（D.Patterson, G.Gibson, R.Katz, "A Case for Redundant Arrays of Inexpensive Disks (RAID)", ACM SIGMOD conference proceedings, 1988, pp. 109-116）」に述べられている。

【0003】RAID3とは、ディスクアレイを構成するディスクをデータを格納するディスク（以下、データディスクと呼ぶ）と冗長データを格納するディスク（以下、パリティディスクと呼ぶ）とに分け、ホストコンピュータから見える論理的なディスクのアドレスが、実際

3

にデータを格納する物理ディスク間を切り替わる大きさ（以下、データストライプサイズと呼ぶ）を、１ビットないし１バイト、あるいは１セクタ等と小さくとったものである。これにより、典型的なディスクアクセスが、アレイを構成する全てのディスクにまたがるようにしてある。ＲＡＩＤ３は、大量データの一括転送能力が優れている。

【０００４】一方、ＲＡＩＤ４とは、ＲＡＩＤ３と同様に、ディスクアレイを構成するディスクをデータディスクとパリティディスクを分け、データストライプサイズを典型的なディスクアクセスのデータサイズ以上にして、典型的なディスクアクセスが、１台の物理ディスクに収まるようにしたものである。ＲＡＩＤ４では、複数個のライトアクセス間で、パリティディスクが競合するという問題がある。

【０００５】また、ＲＡＩＤ５は、ＲＡＩＤ４におけるパリティディスク競合の問題を解決するために、冗長データをディスクアレイを構成する複数台のディスク間に分散して格納するようにしたものである。ＲＡＩＤ５において、あるディスクが故障したときの、デグレード時の動作や性能、予備のホットスタンバイディスクへの故障ディスクのデータの回復の方法や性能に関しては、例えば、エム・ホランド、ジー・ギブソン、等による「パリティ　デクラスタリング　フォー　コンティニアス　オペレーション　イン　リダンダント　ディスク　アレイズ、エーエスピーロス－ブイ　プロシーディングズ」（M.Holland, G.Gibson, "Parity Declustering for Continuous Operation in Redundant Disk Arrays", ASPLOS-V Proceedings, October 12-15, 1992, pp.2-9.）に開示されている。

【０００６】

【発明が解決しようとする課題】上記従来技術によれば、ディスクアレイ内のディスク故障に対する信頼性は向上されるが、ディスク故障が存在するときに、データのリード／ライトを行なう場合、ディスク故障のない時に比べ、ディスクアクセス回数が増大するという問題があった。

【０００７】例えば、ＲＡＩＤ４のディスクアレイでは、通常、少量のデータの読み出しであれば、ディスクのアクセス回数は１回で済む。しかし、ディスクアレイ内のあるデータディスクが故障したとき、故障したディスク内のデータを読み出そうとすると、正常なデータディスク、及び冗長ディスクから、リードしようとするデータを復元するのに必要なデータおよび冗長データを読み出し、排他的論理和を取るなどして、目的とするデータを復元する必要がある。従って、データの冗長構成の単位であるパリティグループが、ｎ個のデータとｍ個の冗長データにより構成されるものを考えた場合、少なくとも、残りの正常なデータディスクｎ－１台と、冗長ディスク１台から、データを読み出さなければならないことになる。このため、最低でも計ｎ台のディスクをアクセスする必要が生じる。つまり、故障したディスク内のデータを読み出そうとした場合には、ディスクアクセス回数が、最低でも通常のリード時のｎ倍に増加することになる。

4

【０００８】同様に、ＲＡＩＤ５においても、通常、少量のデータの読み出し時には、ディスクのアクセス回数は１回で済む。これに対して、ディスクアレイ内のディスクが故障したとき、故障ディスク内のデータを読み出すためには、データの復元を行うために、残りの正常なディスクのうち、データを格納しているディスクｎ－１台と、冗長データを格納しているディスク少なくとも１台をアクセスして、データの復元に必要な、ｎ－１個のデータと、少なくとも１個の冗長データを、読み出す必要がある。従って、この場合にも、最低ｎ台のディスクをアクセスする必要がある。

【０００９】以上、データの読み出しの場合を例に説明したが、同様のことがデータの書き込み時にも問題となる。

【００１０】従って、本発明の目的は、ディスクアレイに関する上述したような問題を解決し、ディスク故障時における故障ディスク内のデータに対するリード／ライト時の、ディスクアクセス回数の増加を防止したディスクアレイの制御方法を提供することにある。

【００１１】

【課題を解決するための手段】上記目的を達成するために、本発明におけるディスクアレイの制御方法は、データを格納するための少なくとも１のディスク装置と冗長データを格納するための少なくとも１のディスク装置とを備えたディスクアレイシステムにおけるディスクアレイ制御方法において、前記データを格納するディスク装置および冗長データを格納するディスク装置の故障を検知し、前記故障が検知された場合、前記冗長データを格納するディスク装置から任意のディスク装置を選択し、前記故障が検知された以外のディスク装置に格納されたデータおよび冗長データを用い前記故障が検知されたディスク装置のデータを復元し、該復元されたデータを前記選択により選択されたディスクに記録することを特徴としている。

【００１２】より詳しくは、多重度ｎ（≧１）のデータに対してｍ（≧１）個の冗長データを記録する、ディスクの冗長構成を採用したディスクアレイにおいて、ｋ（ｍ≧ｋ≧１）台のディスクが故障した時に、まず、ｍ個の冗長データのうち、ｋ個の冗長データを選択し、故障したｋ台のディスク内に記録されていたｋ個のデータまたは冗長データのうち、データと、前記ステップ１で選択しなかった冗長データを復元し、正常なディスク内の、前記ステップ１で選択したｋ個の冗長データの格納されていた領域に書き込む。

【００１３】

5

【作用】本発明によれば、ディスクアレイを構成するディスクの一部に故障が生じた際、冗長データを記録したディスクから故障が生じたディスクに格納されたデータまたは冗長データを重ね書きするディスクを選択し、この選択されたディスクに故障により失われたデータを復元して記録する。このため、ホストコンピュータからのリード／ライト要求に対し、ディスク故障が存在する場合であっても復元されたデータを用いて処理を行なうことができ、アクセス回数の増大を抑えることが可能である。従って、ディスク故障時における性能低下を防止することができる。

【００１４】

【実施例】以下、本発明につき、図面を用い詳細に説明する。

【００１５】図１は、本発明が適用される計算機システムの構成図であり、計算機システム内におけるディスクアレイの位置付けを簡略に示したものである。図１において、１００はディスクアレイであり、ディスクアレイ内においてデータのリード／ライト等の制御を行なうディスクアレイコントローラ２００と、複数のディスクドライブ３００から構成されている。４００は、ディスクアレイ１００が接続されるホストコンピュータである。このような計算機システムにおいて、ディスクアレイ１００は、複数台のディスクドライブ３００をホストコンピュータ４００に論理的に一つのディスク装置に見せる機能を持つ。以下、本実施例では、ＲＡＩＤ５構成のディスクアレイについて述べるが、本発明は、ＲＡＩＤ４構成のディスクアレイにも適用できることはいうまでもない。

【００１６】図２は、ホストコンピュータから見たディスクアレイのイメージである論理ディスクと、実際にデータを格納する物理ディスクとの間の対応を示したマッピング機能図である。

【００１７】図２において、２００はディスクアレイコントローラ、３０１はホストコンピュータから見えるディスクアレイのイメージである論理ディスク、３００は実際にデータを格納するディスクドライブ、即ち物理ディスクである。５０１は論理ディスクにおけるデータの管理単位であるデータブロック、５００、６００は物理ディスクにおけるデータの管理単位であり、それぞれデータを格納するデータブロック、冗長データを格納するパリティブロックである。図２は、１台の論理ディスクがｎ＋ｍ台の物理ディスクによって実現される構成を示している。ｎ＋ｍ台の物理ディスクを、ＤＩＳＫ＃１，ＤＩＳＫ＃２，，，ＤＩＳＫ＃ｎ，ＤＩＳＫ＃ｎ＋１，，，ＤＩＳＫ＃ｎ＋ｍと表現する。ここで、論理ディスク３０１のデータブロック５０１、物理ディスクのデータブロック５００、パリティブロック６００は全て同一の大きさであるとする。この大きさは、論理ディスク３０１内のデータが、物理ディスク３００間を切り替わるデータの折返し単位（ストライプサイズと呼ぶ）であるとともに、冗長データが、物理ディスク間を切り替わる冗長データのストライプサイズでもある。

6

【００１８】図２に示したアレイ構成は、ＲＡＩＤ５のレフトシンメトリック（left symmetric）と呼ばれる構成である。冗長データはｎ個のデータに対して、ｍ個作成され、ｎ多重のデータに対して、冗長度がｍの冗長構成となる。７００はデータの冗長構成の単位であるパリティグループであり、ｎ個のデータブロックと、それに対応したｍ個のパリティブロックからなる。論理ディスク３０１と物理ディスク３００の対応は次のようになる。論理ディスク３０１内のデータを、先頭からｎ個の連続したデータブロック毎に区切る。このｎ個のデータブロック毎に、ｍ個の冗長データを作成し、ｎ個のデータとｍ個の冗長データで一つのパリティグループ７００を構成する。各パリティグループ７００には、論理ディスク内のアドレスの順に、パリティグループ番号を付ける。また、ｍ個の冗長データには、一連の記号Ｐ１、Ｐ２、…、Ｐｍを付ける。

【００１９】パリティグループ７００内のデータと冗長データの、物理ディスクへの格納は次のように行なわれる。

【００２０】論理ディスク３０１内のデータブロックに対して、その番号をｎの剰余系で考えて、１～ｎの整数値を与える。次に、対応するｍ個の冗長データＰ１、Ｐ２、…、Ｐｍに、それぞれ整数値ｎ＋１、ｎ＋２、…、ｎ＋ｍを与え、パリティグループ内の、ｎ個のデータブロック５００とｍ個のパリティブロック６００に対して、１からｎ＋ｍまでの整数値を与える。次に、パリティグループ番号１のデータブロック５００およびパリティブロック６００を、それらに与えられた整数値ｉと等しいディスク番号を有する物理ディスク３００に格納する。つまり、パリティグループ番号１のパリティグループについては、データブロック１を、ＤＩＳＫ１に格納し、データブロック２、３、…、ｎをそれぞれ、ＤＩＳＫ２、ＤＩＳＫ３、…、ＤＩＳＫｎに格納する。パリティブロックについては、Ｐ１からＰｍまでを、それぞれＤＩＳＫｎ＋１からＤＩＳＫｎ＋ｍに格納することになる。パリティグループ番号ｊ（但し、ｊ≧２）のパリティグループについては、前述の方法によって整数値が与えられた各データブロック５００及びパリティブロック６００を、パリティグループ番号ｊ－１内の同じ整数値が与えられたデータブロック５００およびパリティブロック６００が格納される物理ディスクのディスク番号を、１だけ減じたディスク番号を持つ物理ディスク３００に格納する。即ち、図２に示すようにパリティグループ番号が１増す毎に、各ブロックの格納位置が、左へ１つづつずれてゆくことになる。

【００２１】図３には、ディスクアレイの構成図を示す。図３において、２００はディスクアレイコントロー

7

ラ、３００はディスクドライブ、４００はホストコンピュータである。ディスクアレイコントローラ２００は、ＣＰＵ２１０、冗長データの生成、及び障害ディスク内のデータを復元するパリティ生成・データ復元回路２２０、ホストコンピュータ４００との間のでデータの受渡し等を行なうホストインタフェース２３０、それぞれ対応するディスクドライブを制御するドライブコントローラ２４０、メモリ２６０、及びメモリ２６０へのデータのリード／ライトを制御するメモリコントローラ２５０を含んで構成される。これらディスクアレイコントローラ２００の構成要素は、バス２７０により互いに接続されている。

【００２２】メモリ２６０には、各種の処理をＣＰＵ２１０が実施するためのプログラムと、各ディスクドライブ３００の状態を管理するディスク状態管理テーブル２６１、冗長データを管理するためのパリティ管理テーブル２６２、ディスク故障からの回復処理の状態を示す回復処理管理テーブル２６３、各パリティグループ毎にデータ及び冗長データがどのディスクドライブに格納されているかを示すパリティ構成管理テーブル２６４、および回復処理に用いられ、データが上書きされている冗長データの種別を記録する上書き用パリティ管理テーブル２６５等の制御テーブルが格納されている。

【００２３】以下、本実施例において中心的役割を果たすディスク故障時における冗長構成の再構成の方法について説明する。なお、ここでは、説明の便宜上、６台のディスクで構成され、各パリティグループが４個のデータブロックと２個のパリティブロックからなるディスクアレイを想定して説明する。

【００２４】図４、図５に、ディスク状態管理テーブル２６１の論理的な構成を示す。ディスク状態管理テーブル２６１には、各ディスク毎に、それが正常であるか、故障中であるかが示されている。図４は、全てのディスクが正常状態にある場合のディスク状態管理テーブルを示しており、図５はＤＩＳＫ＃２のディスクが故障している場合のディスク状態管理テーブルを示している。

【００２５】図６、図７にはパリティ管理テーブル２６２の論理的な構成を示す。パリティ管理テーブルは、個々の冗長データがそれぞれ冗長データとして使用中であるか、ディスク故障により故障ディスクのデータを上書きされたものであるか（使用中止）を管理する。図６に示すパリティ管理テーブルは、２つの冗長データがともに冗長データとして使用されていることを示しており、図７に示すパリティ管理テーブルは、冗長データＰ２がデータの上書き用に用いられていることを示している。

【００２６】図８に、回復処理管理テーブル２６３の論理的構成を示す。回復処理管理テーブル２６３は、パリテイグループ毎に、故障したディスク内のデータを冗長データが記録されているディスクに上書きすることにより回復済みか、まだ未回復であるかを管理する。

8

【００２７】図９乃至図１４には、パリティ構成管理テーブル２６４の論理的構成を示す。パリティ構成管理テーブル２６４は、各パリティグループ毎に、各物理ディスク内に格納されているデータ、冗長データを識別する情報が記録される物理論理変換部と、データおよび冗長データが、どの物理ディスクに格納されているかを記録する論理物理変換部とを有している。本実施例ではデータＤ１、Ｄ２、Ｄ３、Ｄ４をそれぞれ識別する情報として整数値１、２、３、４を、また、冗長データＰ１、Ｐ２を識別する情報として、整数値５、６を用いている。なお、図９から図１４に示される各パリティ管理テーブルの意味するところは、以下に説明する本実施例の動作とともに説明する。

【００２８】図１６は、本実施例におけるディスク故障モニタ処理の流れを示すフローチャートである。ディスク故障モニタ処理では、ディスク故障の検知、回復処理の起動などを行う。

【００２９】ディスク故障モニタ処理では、各ディスクの状態を調べ、ディスク故障の検知を行い（ステップ１７０１）、ディスクが故障していなければ、そのまま一定時間待ち、再びステップ１７０１へ戻ってディスク故障の検知を行なう（ステップ１７０８）。ステップ１７０１で、ディスクの故障が検知されると、故障ディスクの特定を行ない（ステップ１７０２）、以降、故障ディスクを制御から切り離して動作するために、制御テーブルの設定を行う（ステップ１７０３）。ここでは制御テーブルの設定として、ディスク状態管理テーブル２６１の設定とパリティ構成管理テーブル２６４の設定を行う。

【００３０】ディスクアレイを構成する６台のディスクが全て正常である場合、ディスク状態管理テーブル２６１、および、パリティ構成管理テーブル２６４は、それぞれ図４、図９に示す状態にある。ここで、ステップ１７０１においてディスク故障が検知され、ステップ１２０２において故障を生じたディスクがＤＩＳＫ＃２であることが特定されたとする。ステップ１７０３では、まず、図５に示すようにディスク管理テーブル２６１のＤＩＳＫ＃２の状態を、正常を示す状態から故障を示す状態に変更する。次に、図９に示す正常時のパリティ構成管理テーブル２６４から、ディスク故障時のリード／ライト制御に使用する、故障時のパリティ構成管理テーブル２６１を作成する。具体的には、正常時のパリティ構成管理テーブルの物理論理変換部の物理ディスク番号２の列内のデータを無効とする。更に、論理物理変換部の、物理ディスク番号２のデータを格納していたエントリ内のデータを無効にする。このようにして作成したパリティ構成管理テーブルを図１０に示す。

【００３１】制御テーブルの設定が終了すると、ディスクが故障する前に使用していた冗長データの個数、即ち、ディスク故障検知前の冗長度が１以上であったかど

9

うかチェックする（ステップ１７０４）。このチェックは、パリティ管理テーブル２６２を調べることにより行なうことができる。パリティ管理テーブル２６２内に、使用中の冗長データがｕ個あれば、冗長度ｕで動作していたことになる。上長度検査の結果、ディスク故障検知前に冗長度０で動作していた場合は、ディスク故障により、以降の動作が不可能となるので、ホストコンピュータに障害を報告し、停止する（ステップ１７０９）。ディスク故障検知前に冗長度１以上で動作していた場合は、ステップ１７０５へ進み、故障ディスクのデータを上書きする冗長データを、パリティ管理テーブル２６２から選択し、パリティ管理テーブル２６２にを更新する。ここでは、２つの冗長データのうち、２番目の冗長データであるＰ２を故障ディスクのデータ上書き用に選択し、パリティ管理テーブルのＰ２の状態情報を使用中から上書き用に変更する。変更後のパリティ管理テーブル２６２を図７に示す。また、選択した冗長データの種別、Ｐ２を上書き用パリティ管理テーブル２６５に記録する。

【００３２】次に、回復処理管理テーブル２６３の、全てのパリティグループ番号の回復状態管理情報を未回復に設定する。また、回復済みのパリティグループに対するリード／ライト処理を制御するために、図１１に示すような、データ上書き後のパリティ構成管理テーブルを作成する（ステップ１７０６）。それから、非同期の回復処理を起動し（ステップ１７０７）、回復処理終了後、ステップ１７０１へ戻り、ディスク故障のモニタリングを続ける。

【００３３】次に、ステップ１７０７で実施される回復処理を説明する。

【００３４】図１７は回復処理の流れを示すフローチャートである。

【００３５】図１７において、ｋは処理中のパリティグループの番号を保持する内部変数である。回復処理は、ディスクアレイに対するホストコンピュータからのリード／ライトとは非同期に行なわれる。また、回復処理は、パリティグループ番号１からパリティグループ番号順に行なわれる。

【００３６】回復処理では、まず、内部変数ｋの値を１にする（ステップ１８０１）。次に、内部変数ｋが、ディスクアレイ内の最大のパリティグループ番号の値、ＰＧＮmaxより大きいか否かを判定する（ステップ１８０２）。内部変数ｋがＰＧＮmaxより大であれば、全てのパリティグループについて回復処理を完了したことになるので、処理を終了する。内部変数ｋがＰＧＮmax以下であれば、パリティグループ番号ｋのパリティグループについて回復処理を行う（ステップ１８０３）。パリティグループ番号ｋのパリティグループについて回復処理終了後、内部変数ｋに１を加え、次のパリティグループの処理へ進む（ステップ１８０４）。

10

【００３７】図１８は、図１７に示す回復処理１８０３の詳細な流れを示すフローチャートである。

【００３８】まず、処理対象のパリティグループ（パリティグループ番号ｋ）が、回復済みか否かを、回復処理管理テーブル２６３により調べ、処理対象のパリティグループ回復済みであれば、回復処理を終了する（ステップ１９０１）。

【００３９】一方、処理対象のパリティグループが未回復であれば、そのパリティグループ内で、故障ディスクに格納されていたデータ／冗長データの種別ｃを求める（ステップ１９０２）。これは次の方法によって可能である。まず、処理対象のパリティグループ番号の値を、データブロック及びパリティブロックの総数ｎ＋ｍ（本実施例では６）で除してやり、余りが０であればｎ＋ｍを、余りが０以外であればその余りを、そのパリティグループのパリティレイアウト番号として求める。ここで、パリティレイアウトとは、パリティグループにおいて、各物理ディスクに、どのようにデータおよび冗長データを記録させるかという、データ及び冗長データの物理ディスクへの配置パターンをいう。多重度ｎのデータに、ｍ個の冗長データを付加する冗長構成のＲＡＩＤ５では、パリティレイアウトは、計（ｎ＋ｍ）個存在する。次に、図９に示す正常時のパリティ構成管理テーブル２６４を用い、該当するパリティレイアウト番号の列と、物理論理変換部の故障ディスクの物理ディスク番号の行で特定されるエントリ内の値を求める。これが、データ／冗長データの種別ｃである。

【００４０】次にステップ１９０２で求めた種別ｃが、上書きに用いられる冗長データか否かを上書き用パリティ管理テーブル２６５を参照して判定する（ステップ１９０３）。種別ｃが、上書きに用いられる冗長データであれば、上書き処理は不要であり、ステップ１９０８へジャンプする。種別ｃが上書きに用いられる冗長データでなければ、ステップ１９０４へ進み、種別ｃがデータか冗長データかを判定する。これは、ステップ１９０３同様、正常時のパリティ構成管理テーブル２６４を参照することにより可能である。種別ｃがデータであれば、そのデータを回復するために、同一のパリティグループに属するデータと冗長データを格納しているディスクの番号を、パリティ構成管理テーブルから求め、それらを該当するディスクから読み出し（ステップ１９０５）、故障したディスク内のデータを回復する（ステップ１９０６）。種別ｃが冗長データであれば、その冗長データを生成するため、パリティグループ内のデータを、ディスクから読み出し（ステップ１９０９）、それらを基に冗長データを生成する（１９１０）。

【００４１】最後に、回復したデータまたは生成した冗長データを、上書き用に選択した冗長データの上に上書きし（ステップ１９０７）、回復状態管理テーブル２６３の処理対象のパリティグループのエントリに対応する

回復状態管理情報を未回復から回復済みへ変更して処理を終了する（ステップ１９０８）。

【００４２】図１９は、回復処理中の、ホストコンピュータから要求に対するリード／ライト処理の流れを示すフローチャートである。

【００４３】ディスクアレイコントローラ２００は、ホストコンピュータからのリード／ライト要求を受けると、まず、リード／ライト要求のあったデータの範囲に対応するパリティグループの範囲（ａ－ｂ）を求める（ステップ２００１）。ここで、ａをリード／ライト範囲に対応する最小のパリティグループ番号とし、ｂを同様に最大のパリティグループ番号とする。パリティグループ番号ａ、ｂは、リード／ライト範囲の論理アドレスの最小値と最大値をそれぞれ、パリティグループ内のデータの多重度ｎとストライプサイズｓとの積ｓｎで割った商として求められる。

【００４４】次に、リード／ライト処理中のパリティグループの番号を保持する内部変数ｌに値ａを代入する（ステップ２００２）。内部変数ｌの値がｂより大か否かを判定し（ステップ２００３）。内部変数ｌがｂより大であれば、リード／ライト範囲に対応するパリティグループの処理が完了しているので、処理を終了する。内部変数ｌがｂ以下であれば、パリティグループｌ内のリード／ライト処理を行う（ステップ２００４）。次に、内部変数ｌに１加算し、ステップ２００３へジャンプして処理を続行する（ステップ２００５）。

【００４５】図２０、および２１に、ステップ２００４の各パリティグループ毎のリード／ライト処理の流れを示す。

【００４６】パリティグループ毎のリード／ライト処理では、まず、処理対象のパリティグループが、ディスク故障から回復済みか否かを、回復処理管理テーブル２６３を用いて判定する（ステップ２１０１）。回復済みであれば、回復済みのリード／ライト処理を行って処理を終了する（ステップ２１０７）。なお、回復済みのリード／ライト処理は、図１１に示すようなデータ上書き後のパリティ構成管理テーブルを参照して、リード／ライトの対象となるデータの格納されたディスクを調べ、そのディスクに対するアクセスを行なうことにより処理を行うことができる。この時のリード／ライト処理は、データ多重度ｎに対し、冗長データの数がｍ－１個の、故障したディスクのない、ディスクアレイに対するリード／ライト処理に帰着するので、詳しい説明は省略する。

【００４７】処理対象のパリティグループが、未回復である時は、ホストコンピュータから要求された処理がリード処理であるかライト処理であるかにより分岐して処理を行う（ステップ２１０２）。

【００４８】ホストコンピュータからの処理の要求がリード処理であるときには、まず処理対象のパリティグループ内のリード領域が、どの物理ディスクに対応するかを求める（ステップ２１０３）。次に、リード領域に故障ディスクを含むか否かチェックする（ステップ２１０４）。リード領域に故障ディスクを含まない場合、ステップ２１０３で求めた、物理ディスクから、データを読み出し（ステップ２１０５）、データをホストコンピュータに転送して処理を終了する（ステップ２１０６）。

【００４９】ステップ２１０４で、リード領域に故障ディスクを含むことがわかったときは、パリティグループ内の、故障ディスクに格納されるデータ以外のデータ全てと、故障ディスク内のデータを回復するのに必要な冗長データを一つ、正常なディスクから読み出す（ステップ２１０８）。そして、故障ディスクのデータを回復してステップ２１０６へ進み（ステップ２１０９）、先と同様に、リード対象のデータをホストコンピュータへ転送して処理を終了する。ここで、ステップ２１０７において読み出すデータおよび冗長データが、どの物理ディスクに格納されているかは、正常時のパリティ構成管理テーブル２６４から求めることができる。

【００５０】ステップ２１０２で、ホストコンピュータからの処理の要求がライト処理と判定された場合には、図２１のラベルＷへジャンプし、処理を続ける。以下、図２１を参照してライト処理につき説明する。

【００５１】まず、正常時のパリティ構成管理テーブル２６４を用いてパリティグループ内のライト領域が、どの物理ディスクに対応するかを求める（ステップ２２０１）。次に、ライト領域に故障ディスクを含むか否かを判定する（ステップ２２０２）。ライト領域に故障ディスクを含まない場合、ステップ２２０１で求めた物理ディスクから、ライト領域に対応する更新前のデータ（旧データ）、および更新前の冗長データを読みだす（ステップ２２０３）。次に、ホストコンピュータから受け取ったライトデータと、ディスクからリードした、旧データ、および更新前の冗長データから、新しい冗長データを生成する（ステップ２２０４）。次に、新しい冗長データと、ライトデータを、それぞれ、該当する物理ディスクに書き込み処理を終了する（ステップ２２０５）。

【００５２】以上により、処理対象のパリティグループ内のライトデータのディスクへの書き込みと、冗長データの更新が完了する。

【００５３】ステップ２２０２で、ライト領域に故障ディスクを含んでいた場合、処理対象のパリティグループ内の物理ディスクでデータを格納しているもののうち、ライト領域に対応する物理ディスクを除いた、残りの物理ディスクからデータをリードする（２２０６）。次に、ライトデータと、ステップ２２０６でリードしたデータから、新しい冗長データを生成する（２２０７）。それから、故障ディスクを除いて、ライトデータと、ステップ２２０７で生成した新しい冗長データを物理ディスクにライトして処理を終了する（ステップ２２０８）。

13

【００５４】以上説明したように、本実施例によれば、回復処理中にも、ホストコンピュータからのリード／ライトアクセスに対する処理を、回復処理と並行して行うことができる。なお、以上の説明では、６台のディスクによりデータの多重度４、冗長度２の冗長構成を採っているときに、２番のディスク（ＤＩＳＫ＃２）が故障したとき例に説明したが、その後さらに、他のディスクが故障した場合にも同様の処理手順により回復処理を行なうことができる。上述したＤＩＳＫ＃２の故障に対する回復処理完了後に、更にＤＩＳＫ＃４が故障した場合を例に、このとき用いられるパリティ構成管理テーブルの例を図１２、図１３、図１４に示す。ＤＩＳＫ＃２の故障に対する回復処理完了後に、更にＤＩＳＫ＃４が故障したときは、残りの冗長データＰ１にＤＩＳＫ＃４に格納されていたデータを上書きすることにより、これを回復することが可能である。

【００５５】以上述べたように、本実施例によれば、ディスク故障時に、ディスクアレイの冗長構成を、冗長度を１だけ落とした、故障ディスクの無い冗長構成に再構成するため、故障ディスク内のデータに対するリード／ライト処理におけるディスクアクセス回数を従来に比べ、低減することができる。また、冗長度が１だけ減るので、ライト時に更新すべき冗長データの数が１だけ少なくなり、ライト時のライトペナルティを減らすこともできる。

【００５６】また、本実施例におけるディスクアレイの制御方法は、ＲＡＩＤ３構成のディスクアレイに適用することも可能である。例えば、多重度ｎのデータに１個の冗長データを付加する冗長構成のＲＡＩＤ３に本実施例における制御方法を適用した場合を考えてみる。データを格納するｎ台のディスクのうち、１台が故障したとする。このとき、データのリード処理は、データをｎ－１台のデータディスクから、また、冗長データをそれを格納する１台のパリティディスクからそれぞれ読み出し、故障したディスクのデータを復元してそのデータをホストコンピュータに送ることにより行なわれる。また、データのライト処理は、ホストコンピュータから送られてきたデータから、冗長データを作成し、故障していないｎ－１台のデータディスクと、１台のパリティディスクに、それぞれデータと冗長データを書き込むことにより行なわれる。この場合、リード処理時には故障したディスク内のデータの復元が不要となり、ライト時には冗長データの生成が不要になるという利点がある。

【００５７】以上説明した実施例では、故障したディスク内のデータが通常データ、および冗長データの何れの場合にもデータ回復を行なっているが、以下、第２の実施例として、故障したディスク内のデータが冗長データである場合にはその回復処理を行なわないものについて説明する。

【００５８】本実施例においては、システム構成等は第１の実施例と同じであり、その説明は省略し、ディスク故障検出時の処理のうち、第１の実施例とは異なる部分について説明する。

14

【００５９】本実施例では、図１７に示す回復処理のうち、ステップ１８０３におけるパリティグループ毎の回復処理が第１の実施例とは多少異なる。図２２に本実施例によるパリティグループ毎の回復処理のフローチャートを示す。なお、図２２では、図１８と同様の処理が行なわれる部分については、図１８と同一の符号を用いている。

【００６０】第１の実施例においても説明したように、まず、処理対象となっているパリティグループが既に回復処理が行なわれたものかどうか判定し（ステップ１９０１）、回復済みであれば処理を終了し、未回復の場合、次のステップ１９０２に進み、処理対象のパリティグループのデータの種別を求める。処理対象のパリティグループ内のデータの種別が求まると、それに続いて、求めた種別が冗長データを示している場合には、そのままステップ１９０８に進んで回復状態管理情報の設定を行ない本処理を終了する。また、ステップ１９０２で求めた種別が冗長データでない場合には、ステップステップ１９０５へ進み、ステップ１９０６、およびステップ１９０７、を実行してデータを回復し、ステップ１９０８で回復状態管理情報の設定を行なって処理を終了する。

【００６１】本実施例における回復処理完了後の、パリティ構成管理テーブルを図１５に示す。ライト時の冗長データ更新、また、更に他のディスク故障が生じた場合のデータ回復には、図１５に示す、パリティ構成管理テーブルを使用する。図１６に示すように、本実施例では冗長データに対する回復処理を行なわないため、故障ディスク内に冗長データを記録していたパリティグループに対応するパリティレイアウトに対しては、上書きすべく選択した冗長データが残っている。即ち、パリティグループによって、保持される冗長データの種別が異なる場合が生じる。このため、ライト時の冗長データ更新や、更にディスクが故障した時のデータ復元処理が若干複雑になってしまう。しかし、故障ディスク内の冗長データについて回復処理を行なわないことにより、ディスクアレイの冗長構成の再構成に要する回復処理の時間、処理データ量を、削減できるというメリットを有している。第１及び第２の実施例では、故障ディスク内のデータ（または冗長データ）の回復処理をホストコンピュータからのディスクアクセスとは非同期にバックグラウンド処理として行なうものについて説明した。次に、第３の実施例として、ホストコンピュータから、故障したディスク内のデータに、リードまたはライトアクセスがあった時、要求データを対応する上書き用に選択された冗長データ上に復元する、オンデマンドでのデータ回復の方法について説明する。

15

【0062】図23、図24を用いて、ホストコンピュータからのリード／ライト要求の処理に同期して回復処理を行う場合の、パリティグループ毎のリード／ライト処理の方法を示す。なお、本実施例においても他の部分は第1の実施例と同様であるので、その説明は省略する。

【0063】ホストコンピュータからのリード／ライト要求があったときには、まず、処理対象のデータを含むパリティグループが回復済みか否かをチェックする（ステップ2301）。このパリティグループが回復済みである場合には、回復済みのリード／ライト処理を行う（ステップ2307）。ここでは、図20に示したリード／ライト処理におけるステップ2107と同様であり、説明は省略する。

【0064】処理対象のパリティグループが未回復であった場合、ホストコンピュータからの処理要求がリードかライトかによって異なる分岐先に進み処理を行う（ステップ2302）。ホストコンピュータからの処理要求がリードのときには、まず、処理対象のパリティグループ内のリード領域が格納される物理ディスクを求める（ステップ2303）。次に、求められた物理ディスクに故障ディスクを含むか否かチェックする（ステップ2304）。故障ディスクを含まない場合には、ステップ2303で求めた物理ディスクからデータを読み出し（ステップ2305）、そのデータをホストコンピュータに転送して処理を終える。（ステップ2306）

ステップ2304で、リード領域に故障ディスクを含むことがわかったときは、そのパリティグループに属し、故障ディスク以外のディスクに格納されているデータ全てと、故障ディスク内のデータを回復するのに必要な冗長データを読み出し（ステップ2308）、それら読み出したデータを用いて故障ディスクに格納されていたデータを回復する（ステップ2309）。その後、ステップ2309で回復した故障ディスクのデータをディスク故障モニタで選択して、上書き用パリティ管理テーブルに記録されている冗長データの上に上書きする（ステップ2310）。さらに、、回復処理管理テーブル内の処理対象のパリティグループに対応する回復状態管理情報に回復済みをセットし（ステップ2311）、ステップ2306へ進んでリード対象のデータをホストコンピュータへ転送する。

【0065】ステップ2302で、ホストコンピュータからの処理要求がライトと判定された場合には、図24のラベルWへジャンプして処理を続ける。

【0066】まず、図9に示す正常時のパリティ構成管理テーブルを用いて、処理対象のデータが属するパリティグループ内のライト領域が、どの物理ディスクに対応しているか求める（ステップ2401）。次にライト領域に故障ディスクを含むか否かを判定する（ステップ2402）。ライト領域に故障ディスクを含まない場合には、ステップ2401で求めた物理ディスクから、ライト領域に対応する更新前のデータ（旧データ）、および更新前の冗長データを読みだす（ステップ2403）。次に、ホストコンピュータから受け取ったライトデータと、ディスクから読み出したライト処理前の旧データと、更新前の冗長データから新しい冗長データを生成する（ステップ2404）。次に、新しい冗長データとライトデータをそれぞれ物理ディスクに書き込んで処理を終了する（ステップ2405）。以上により、処理対象のパリティグループ内のライトデータのディスクへの書き込みと、パリティの更新が完了する。

16

【0067】一方、ステップ2402において、ライト領域に故障ディスクを含むことが分かった場合には、処理対象のデータが属するパリティグループ内の物理ディスクでデータを格納しているもののうち、ライト領域に対応する物理ディスクを除いた、残りの物理ディスクからデータを読み出す（2406）。次に、ライトデータと、ステップ2406で読み出したデータから新しい冗長データを生成する（2407）。それから、故障ディスクを除いた健在なディスクのライトデータをそれぞれのディスクに書き込み、ステップ2407で生成した新しい冗長データのうち、故障ディスクのデータを上書きする冗長データを除く冗長データをそれぞれ対応するディスクにライトし、故障ディスクへのライトデータをディスク故障モニタにより選択された上書き用の冗長データの上に書き込み（ステップ2408）、回復状態管理情報をセットして処理を終了する（ステップ2409）。以上、述べた処理方法により、ホストコンピュータからのリード／ライト要求の処理に同期した、故障ディスク内のデータの正常なディスク内の冗長データ上への上書き処理（回復処理）を行うことができる。なお、本実施例によるオンデマンドでのデータ回復の方法では、故障ディスク内のデータのみを復元し、故障ディスク内の冗長データは復元していない。

【0068】以上説明した第3の実施例では、故障ディスク内のデータの回復処理をホストコンピュータからリード／ライト要求に同期して行なうため、ホストコンピュータからリード／ライト要求の処理に際して復元されたデータを用いることができ、バックグラウンドで回復処理を行なう場合に比べ、回復処理自体でのディスクアクセス回数を低減させることができる。また、本実施例による回復処理を、例えば、第1の実施例で説明したようなバックグラウンドの回復処理と並行して行なうこともできる。この場合には、バックグラウンドの回復処理時に、回復状態管理テーブルを、パリティグループ番号が小さいものから検索して、最もパリティグループ番号が小さい未回復のパリティグループを見つけて回復処理を行えばよい。

【0069】第3の実施例では、ホストコンピュータからのリード／ライト要求に同期して、故障ディスク内の

データにリードまたはライトが直接行われた場合にのみ故障ディスク内のアクセス対象となったデータを回復していたが、リード／ライトの対象となるデータが属するパリティグループ内の故障ディスクに格納されたデータ及び冗長データを上書きすべく選択された冗長データ上に回復してもよい。以下、第４の実施例として、このようなデータの回復方法について、図２５及び図２６を用い説明する。

【００７０】第３の実施例と同様に、まず、ステップ２３０１で、処理対象のパリティグループが回復済みか否かをチェックし、回復済みである場合には、回復済みのリード／ライト処理を行う（ステップ２３０７）。処理対象のパリティグループが未回復であった場合には、ホストコンピュータからの処理要求がリードかライトかによって分岐して処理を行う（ステップ２３０２）。ホストコンピュータからの処理要求がリードの時には、まず処理対象のパリティグループ内のリード領域が、どの物理ディスクに対応するか（ここでは、物理ディスクｄ～ｅが該当するものとする）を求める（ステップ２３０３）。次に、当パリティグループの故障ディスク内のデータ及び冗長データの種別ｆを判定する（ステップ２６０１）。種別ｆが上書き用に選択された冗長データである場合は、物理ディスクｄ～ｅからデータをリードし（ステップ２６０２）、回復処理管理テーブルの回復状態管理情報を回復済みにセットする（２６０３）。さらに、データをホストコンピュータに転送して処理を終える（ステップ２３０６）。

【００７１】ステップ２６０１で種別ｆが上書き用に選択されていない冗長データである場合には、当パリティグループ内の全てのデータをディスクから読み出し（ステップ２６０４）、故障ディスク内の冗長データを生成し（ステップ２６０５）、ステップ２６０５で生成した故障ディスク内の冗長データを上書き用に選択されている冗長データに上書きしてステップ２６０３へ進む（ステップ２６０６）。

【００７２】また、ステップ２６０１で、種別ｆがデータであった場合は、当パリティグループ内の、故障していないディスクに含まれる全てのデータと、冗長データを一つディスクから読み出し（ステップ２６０７）、故障ディスク内のデータを回復した後（ステップ２６０８）、故障ディスク内のデータを上書き用として選択されている冗長データに上書きしてステップ２６０３へ進む（ステップ２６０９）。

【００７３】ステップ２３０２で、ホストコンピュータからの処理要求がライトであるときには、図２６のラベルWへジャンプする。

【００７４】ステップ２４０１でパリティグループ内のライト領域を物理ディスク番号（先程と同様に、物理ディスクｄ～ｅが該当するものとする）に変換し、ステップ２７０１へ進み、当パリティグループ内の故障ディスク内のデータあるいは冗長データの種別（ｆとする）を判定する。種別ｆが上書き用に選択されている冗長データである場合、物理ディスクｄ～ｅから旧データを読み出し、更に故障していないディスクから当パリティグループの冗長データを読み出す（ステップ２７０２）。次に、ステップ２７０２で読み出した旧データ、旧冗長データと、ホストコンピュータからのライトデータとから新しい冗長データを生成する（ステップ２７０３）。そして、新しい冗長データとライトデータをディスクに書き込み（ステップ２７０４）、回復処理管理テーブルの回復状態管理情報に回復済みをセットして処理を終了する（ステップ２７０５）。

【００７５】ステップ２７０１で種別ｆが上書き用に選択されていない冗長データである場合は、物理ディスクｄ～ｅ以外のディスクから当パリティグループの残りのデータを読み出し（ステップ２７０６）、この読み出したデータとライトデータとから、新しい冗長データを生成し（ステップ２７０７）、ライトデータおよび健全なディスクに書き込まれる冗長データをそれぞれ該当するディスクへ書き込み、ステップ２７０５へ進む（ステップ２７０８）。

【００７６】また、ステップ２７０１で種別ｆがデータである場合は、当パリティグループ内のデータを格納しているディスクのうちディスクｄ～ｅを除くディスクからデータをリードし、更に必要なら冗長データをも読み出し（ステップ２７０９）、ライトデータとステップ２７０９で読み出したデータ、冗長データから新しい冗長データを生成し、更に必要であれば、故障ディスク内のデータを復元する（ステップ２７１０）。そして、故障していない健在なディスクのライトデータをディスクライトし、上書きしない冗長データをディスクライトし、ライトデータまたはステップ２７１０で復元したデータとして得られる、故障ディスク内のデータを、上書き用として選択されている冗長データ上に上書きし（ステップ２７１１）、ステップ２７０５へ進む。

【００７７】以上説明した処理により、ホストコンピュータからのリード／ライトに同期して、リード／ライト対象のデータが属するパリティグループ内のデータまたは冗長データであって、故障ディスクに格納されたデータまたは冗長データを上書き用として選択した冗長データ上に回復することができる。

【００７８】以上、本発明について、４つの実施例を基に説明してきたが、以上説明した実施例によれば、ＲＡＩＤ４、あるいはＲＡＩＤ５によるディスクアレイにおいて、一部のディスクに故障が生じた際、ディスクアレイ内の故障ディスクを排除して、健全なディスクのみでデータ構成を再編成するので、ディスク故障時におけるデータアクセス処理の性能低下を防止することができる。図２７、２８にそれぞれＲＡＩＤ４、ＲＡＩＤ５における本発明と従来方式のアクセス回数につき、表にし

て示す。　また、以上説明してきた実施例では、データ格納用のディスクと冗長データ格納用のディスクからなるディスクアレイについて説明したが、本発明は、さらに、ディスク故障時に故障ディスクの代替用として用いられるホットスタンバイディスクを有するディスクアレイに適用することもできる。ホットスタンバイディスクを有するディスクアレイでは、ディスク故障時に、故障ディスク上のデータや冗長データをホットスタンバイディスク上に復元する。この場合でも、ディスクの故障が重なり、全てのホットスタンバイディスクを代替として使いきっているような状態で、更にディスクの故障が発生したようなときに本発明が有効となる。

【００７９】さらに、本発明をＲＡＩＤ３のディスクアレイに適用した場合には、ホストコンピュータからのデータリード要求に対する処理における故障ディスクのデータを復元する復元処理、あるいはホストコンピュータからのデータライト要求に対する処理における冗長データを生成する処理を省けるといった利点がある。

【００８０】

【発明の効果】本発明によれば、ディスクアレイを構成する一部のディスクに故障が生じたときであっても、故障ディスクに対するリード／ライト時のディスクアクセス回数増加を防ぐことができる。また、このような場合における正常なディスクに対するライト時のディスクアクセス回数についても、低減することが可能となり、ディスク故障時の性能低下を抑えることができる。

【図面の簡単な説明】

【図１】本発明が適用される計算機システムの構成図である。

【図２】論理ディスクと物理ディスクの対応関係を示すマッピング機能図である。

【図３】ディスクアレイの構成図である。

【図４】正常時のディスク状態管理テーブルの論理的構成図である。

【図５】ディスク故障時のディスク状態管理テーブルの論理的構成図である。

【図６】正常時のパリティ管理テーブルの論理的構成図である。

【図７】ディスク故障時のパリティ管理テーブルの論理的構成図である。

【図８】回復処理管理テーブルの論理的構成図である。

【図９】正常時のパリティ構成管理テーブルの論理的構成図である。

【図１０】ディスク故障時のパリティ構成管理テーブルの論理的構成図である。

【図１１】再構成後のパリティ構成管理テーブルの論理的構成図である。

【図１２】ディスク２故障時の回復処理完了後のパリティ構成管理テーブルの論理的構成図である。

【図１３】ディスク２の回復処理完了後、更にディスク４が故障したときのパリティ構成管理テーブルの論理的構成図である。

【図１４】ディスク２、４の回復処理完了後のパリティ構成管理テーブルの論理的構成図である。

【図１５】本発明の第２の実施例における再構成後のパリティ構成管理テーブルの論理的構成図である。

【図１６】ディスク故障モニタの処理フローチャートである。

【図１７】非同期の回復処理のフローチャートである。

【図１８】パリティグループ毎の回復処理のフローチャートである。

【図１９】リード／ライト処理のフローチャートである。

【図２０】パリティグループ毎のリード／ライト処理のフローチャートである。

【図２１】パリティグループ毎のリード／ライト処理のフローチャートである。

【図２２】本発明の第２の実施例によるパリティグループ毎の回復処理のフローチャートである。

【図２３】リード／ライトに同期した回復処理を行う場合のパリティグループ毎のリード／ライト処理のフローチャートである。

【図２４】リード／ライトに同期した回復処理を行う場合のパリティグループ毎のリード／ライト処理のフローチャートである。

【図２５】パリティグループ毎のリード／ライト処理のフローチャートである。

【図２６】パリティグループ毎のリード／ライト処理の続きである。

【図２７】ＲＡＩＤ４における本発明と従来技術との効果の差を示す比較図である。

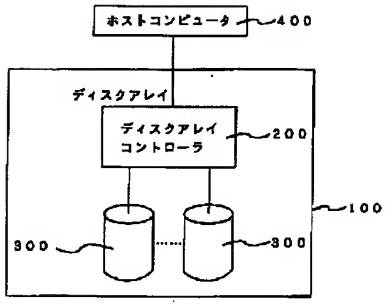
【図２８】ＲＡＩＤ５における本発明の従来技術との効果の差を示す比較図である。

【符号の説明】

１００…ディスクアレイ、２００…ディスクアレイコントローラ、２１０…ＣＰＵ、２２０…パリティ生成・データ復元回路、２３０…ホストインタフェース、２４０…ドライブコントローラ、２５０…メモリコントローラ、２６０…メモリ、２７０…バス、３００…ディスクドライブ、３０１…論理ディスク、４００…ホストコンピュータ、５００，５０１…データブロック、６００…パリティブロック、７００…パリティグループ、７０１…パリティグループ番号である。

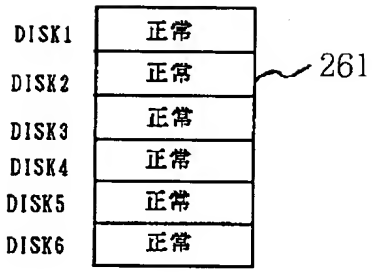
【図1】

図1



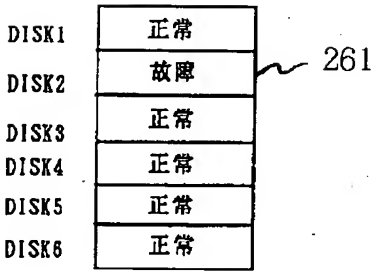
【図4】

図4



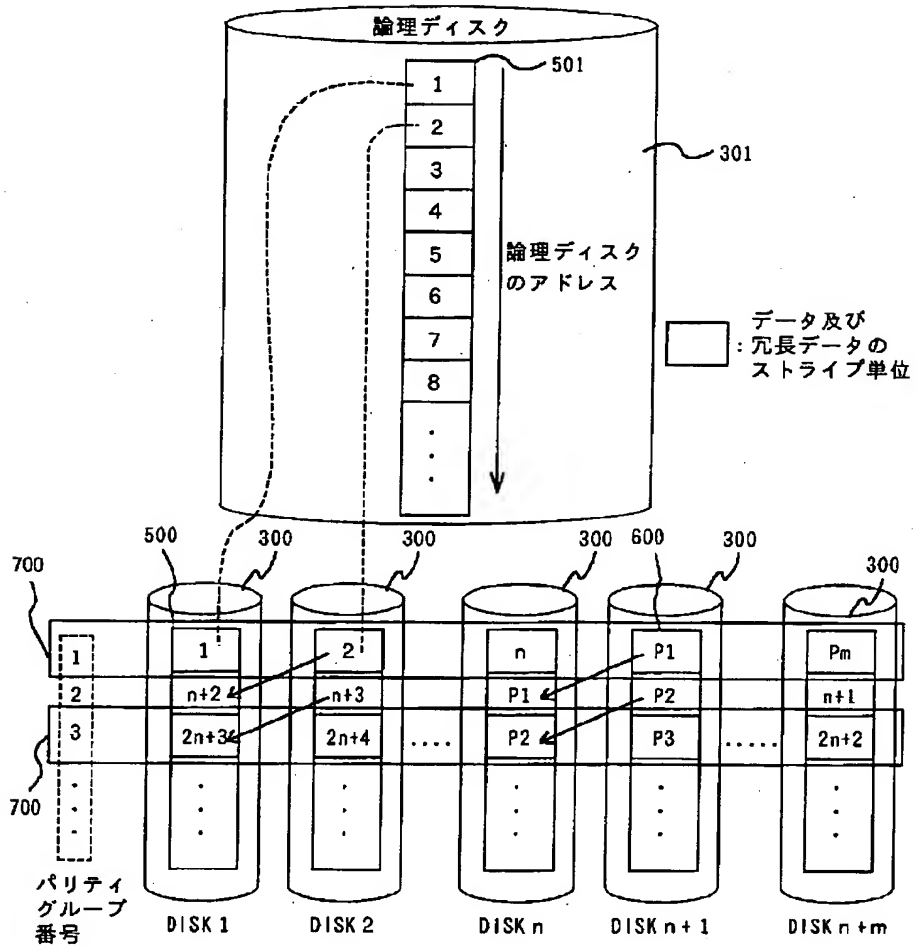
【図5】

図5



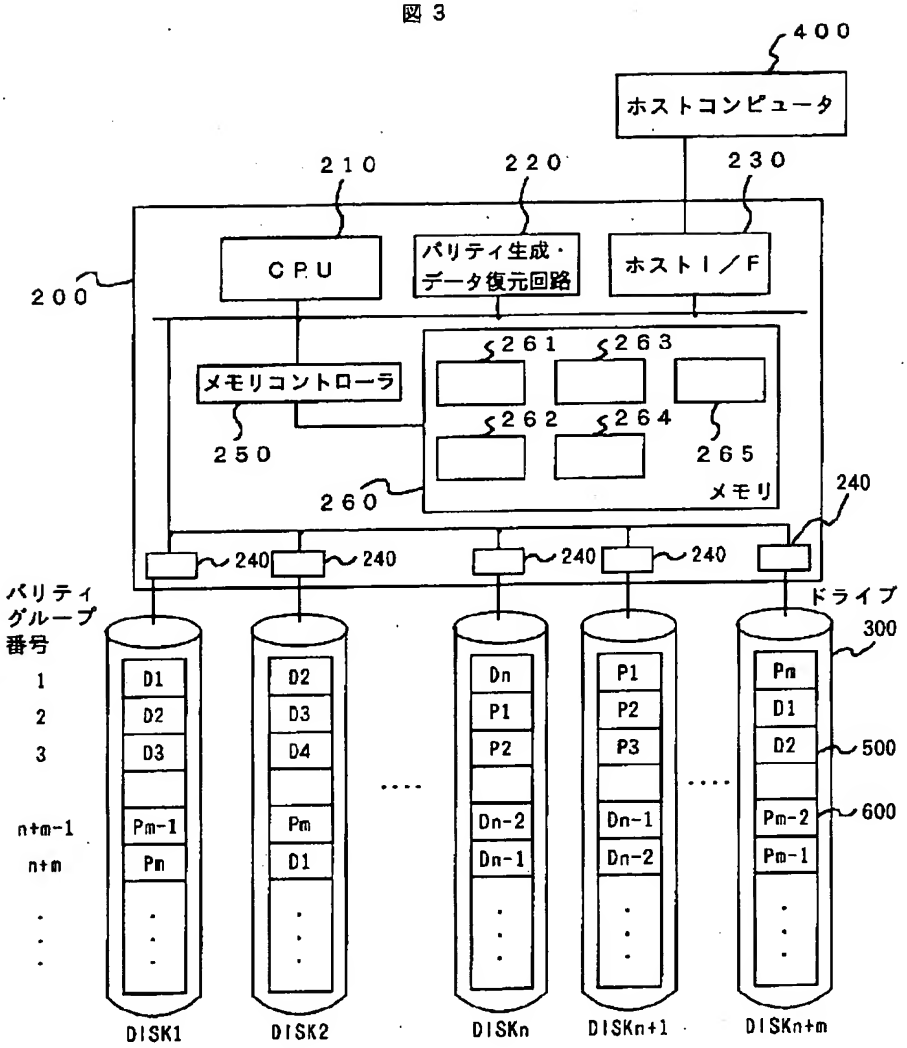
【図2】

図2



【図3】

図 3



【図6】

図 6

パリティ種別	状態情報
P1	使用中
P2	使用中

262

【図7】

図 7

パリティ種別	状態情報
P1	使用中
P2	上書き用

262

【図8】

図8

ﾊﾟﾘﾃｨｸﾞﾙｰﾌﾟ番号	回復状態管理情報
1	回復済み
2	回復済み
3	回復済み
4	回復済み
k	回復済み
k+1	未回復
PGNmax	未回復

263

【図9】

図9

ﾊﾟﾘﾃｨﾚｲｱｳﾄ番号	物理論理変換部						論理物理変換部					
	物理ﾃﾞｨｽｸ番号						ﾃﾞｰﾀ/ﾊﾟﾘﾃｨ番号					
	1	2	3	4	5	6	1	2	3	4	5	6
1	1	2	3	4	5	6	1	2	3	4	5	6
2	2	3	4	5	6	1	6	1	2	3	4	5
3	3	4	5	6	1	2	5	6	1	2	3	4
4	4	5	6	1	2	3	4	5	6	1	2	3
5	5	6	1	2	3	4	3	4	5	6	1	2
6	6	1	2	3	4	5	2	3	4	5	6	1

【図10】

図10

ﾊﾟﾘﾃｨﾚｲｱｳﾄ番号	物理論理変換部						論理物理変換部					
	物理ﾃﾞｨｽｸ番号						ﾃﾞｰﾀ/ﾊﾟﾘﾃｨ番号					
	1	2	3	4	5	6	1	2	3	4	5	6
1	1	・	3	4	5	6	1	・	3	4	5	6
2	2	・	4	5	6	1	6	1	・	3	4	5
3	3	・	5	6	1	2	5	6	1	・	3	4
4	4	・	6	1	2	3	4	5	6	1	・	3
5	5	・	1	2	3	4	3	4	5	6	1	・
6	6	・	2	3	4	5	・	3	4	5	6	1

【図１１】

図１１

ﾊﾟﾘﾃｨﾚｲｱｳﾄ番号	物理論理変換部	論理物理変換部
	物理ﾃﾞｨｽｸ番号	ﾃﾞｰﾀ/ﾊﾟﾘﾃｨ番号
	1 2 3 4 5 6	1 2 3 4 5 6
1	1 ・ 3 4 5 2	1 6 3 4 5 ・
2	2 ・ 4 5 3 1	6 1 5 3 4 ・
3	3 ・ 5 4 1 2	5 6 1 4 3 ・
4	4 ・ 5 1 2 3	4 5 6 1 3 ・
5	5 ・ 1 2 3 4	3 4 5 6 1 ・
6	1 ・ 2 3 4 5	1 3 4 5 6 ・

【図１２】

図１２

ﾊﾟﾘﾃｨﾚｲｱｳﾄ番号	物理論理変換部	論理物理変換部
	物理ﾃﾞｨｽｸ番号	ﾃﾞｰﾀ/ﾊﾟﾘﾃｨ番号
	1 2 3 4 5 6	1 2 3 4 5 6
1	1 ・ 3 4 5 2	1 6 3 4 5 ・
2	2 ・ 4 5 3 1	6 1 5 3 4 ・
3	3 ・ 5 4 1 2	5 6 1 4 3 ・
4	4 ・ 5 1 2 3	4 5 6 1 3 ・
5	5 ・ 1 2 3 4	3 4 5 6 1 ・
6	1 ・ 2 3 4 5	1 3 4 5 6 ・

【図１３】

図１３

ﾊﾟﾘﾃｨﾚｲｱｳﾄ番号	物理論理変換部	論理物理変換部
	物理ﾃﾞｨｽｸ番号	ﾃﾞｰﾀ/ﾊﾟﾘﾃｨ番号
	1 2 3 4 5	6 1 2 3 4 5 6
1	1 ・ 3 ・ 5 2	1 6 3 ・ 5 ・
2	2 ・ 4 ・ 3 1	6 1 5 3 ・ ・
3	3 ・ 5 ・ 1 2	5 6 1 ・ 3 ・
4	4 ・ 5 ・ 2 3	・ 5 6 1 3 ・
5	5 ・ 1 ・ 3 4	3 ・ 5 6 1 ・
6	1 ・ 2 ・ 4 5	1 3 ・ 5 6 ・

【図１４】

図１４

ﾊﾟﾘﾃｨﾚｲｱｳﾄ番号	物理論理変換部						論理物理変換部					
	物理ﾃﾞｨｽｸ番号						ﾃﾞｰﾀ/ﾊﾟﾘﾃｨ番号					
	1	2	3	4	5	6	1	2	3	4	5	6
1	1	・	3	・	4	2	1	6	3	5	・	・
2	2	・	4	・	3	1	6	1	5	3	・	・
3	3	・	4	・	1	2	5	6	1	3	・	・
4	4	・	1	・	2	3	3	5	6	1	・	・
5	2	・	1	・	3	4	3	1	5	6	・	・
6	1	・	2	・	4	3	1	3	6	5	・	・

【図１５】

図１５

ﾊﾟﾘﾃｨﾚｲｱｳﾄ番号	物理論理変換部						論理物理変換部					
	物理ﾃﾞｨｽｸ番号						ﾃﾞｰﾀ/ﾊﾟﾘﾃｨ番号					
	1	2	3	4	5	6	1	2	3	4	5	6
1	1	・	3	4	5	2	1	6	3	4	5	・
2	2	・	4	5	3	1	6	1	5	3	4	・
3	3	・	5	4	1	2	5	6	1	4	3	・
4	4	・	6	1	2	3	4	5	6	1	・	3
5	5	・	1	2	3	4	3	4	5	6	1	・
6	1	・	2	3	4	5	1	3	4	5	6	・

【図２７】

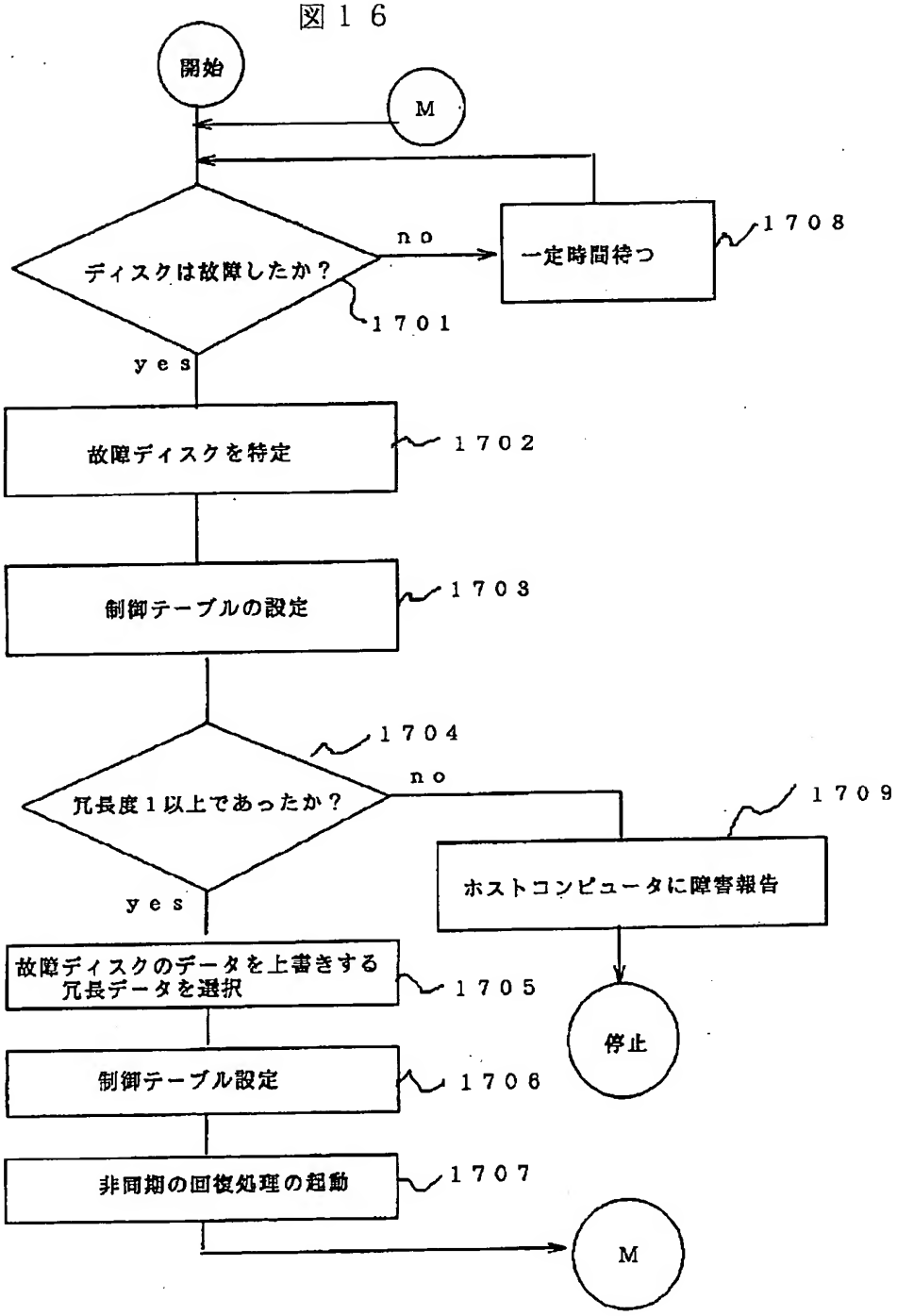
図２７

	ディスクアクセス回数		
	正常時	故障時	本発明
故障ディスクのデータリード		n	1
データライト		n+m-1	2m
正常ディスクのデータリード	1	1	1
データライト	2(m+1)	2(m+1)	2m

（注）データディスクｎ台、パリティディスクｍ台の構成を仮定する。
故障はディスク１台の故障を仮定する。

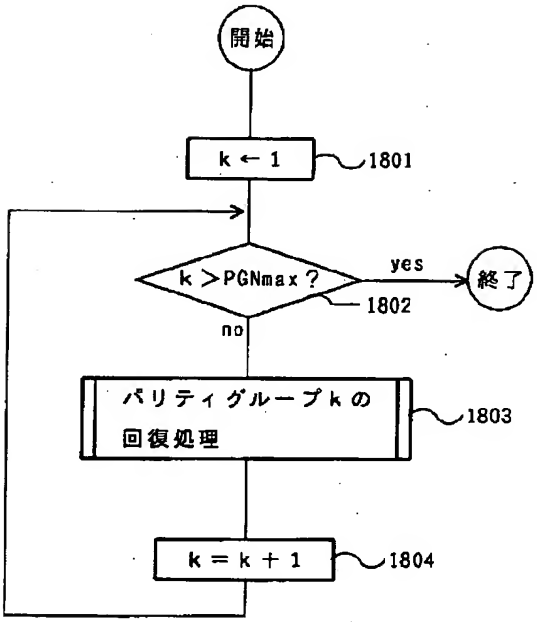
【図16】

図16


開始
M
ディスクは故障したか？
no
1701
一定時間待つ
1708
yes
故障ディスクを特定
1702
制御テーブルの設定
1703
1704
冗長度1以上であったか？
no
1709
ホストコンピュータに障害報告
yes
故障ディスクのデータを上書きする冗長データを選択
1705
停止
制御テーブル設定
1706
非同期の回復処理の起動
1707
M

【図17】

図17



【図28】

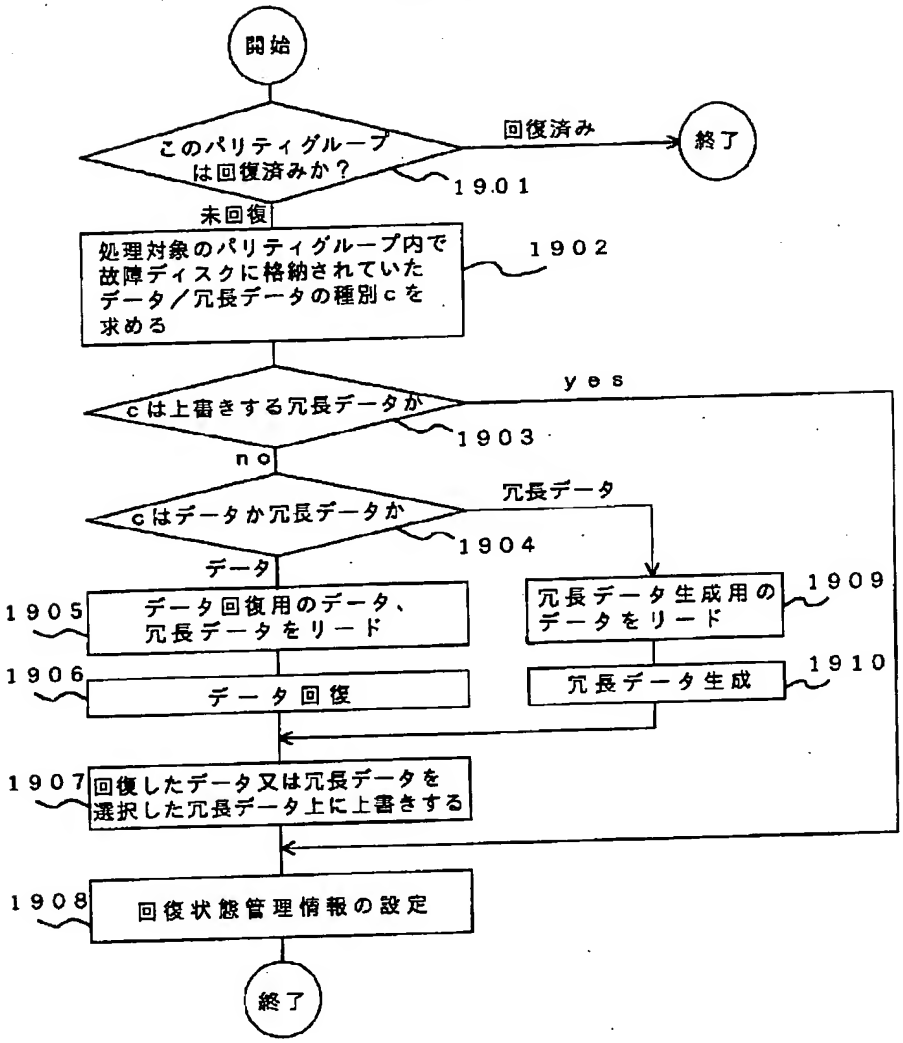
図28

		ディスクアクセス回数		
		正常時	故障時	本発明
故障ディスク内のデータリード			n	1
データライト			n+m-1	2m
対応冗長ﾃﾞｰﾀが	ﾃﾞｰﾀﾘｰﾄﾞ		1	1
故障ﾃﾞｨｽｸ内にある	ﾃﾞｰﾀﾗｲﾄ		2m	2m
対応冗長ﾃﾞｰﾀが	ﾃﾞｰﾀﾘｰﾄﾞ	1	1	1
正常ﾃﾞｨｽｸ内にある	ﾃﾞｰﾀﾗｲﾄ	2(m+1)	2(m+1)	2m

（注）パリティグループとして、（n＋m）台のディスクを用いて
データをn台のディスクに格納し、冗長データをm台の
ディスクに格納するものを仮定する。
故障はディスク1台の故障を仮定する。

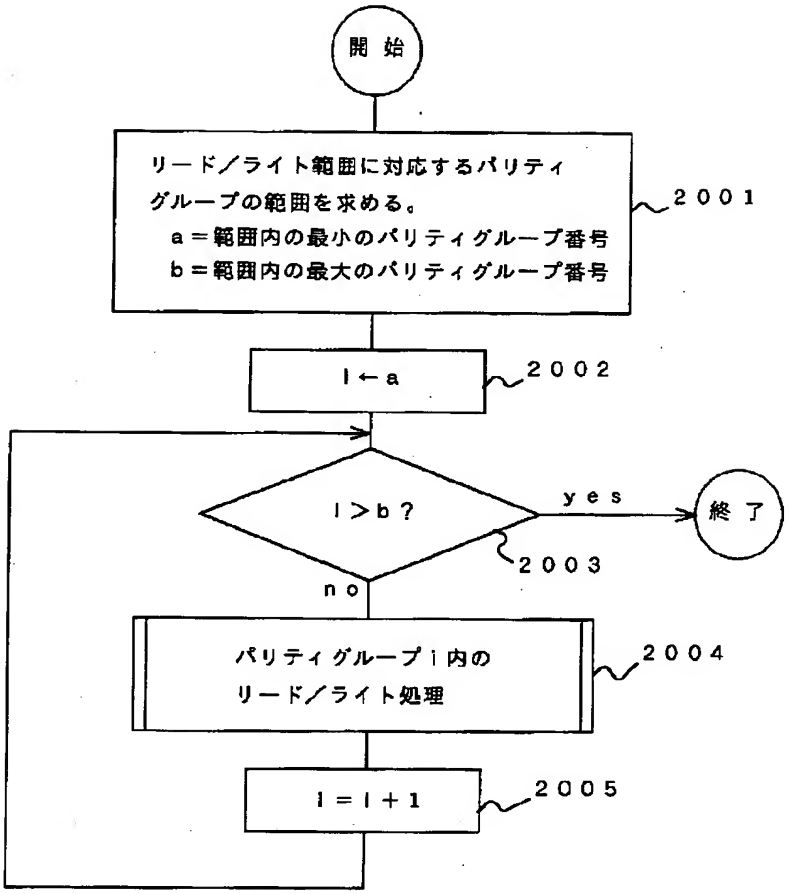
【図18】

図18


開始
このパリティグループは回復済みか？
回復済み
終了
1901
未回復
処理対象のパリティグループ内で故障ディスクに格納されていたデータ／冗長データの種別cを求める
1902
cは上書きする冗長データか
yes
1903
no
cはデータか冗長データか
冗長データ
1904
データ
1905
データ回復用のデータ、冗長データをリード
1906
データ回復
冗長データ生成用のデータをリード
1909
冗長データ生成
1910
1907
回復したデータ又は冗長データを選択した冗長データ上に上書きする
1908
回復状態管理情報の設定
終了

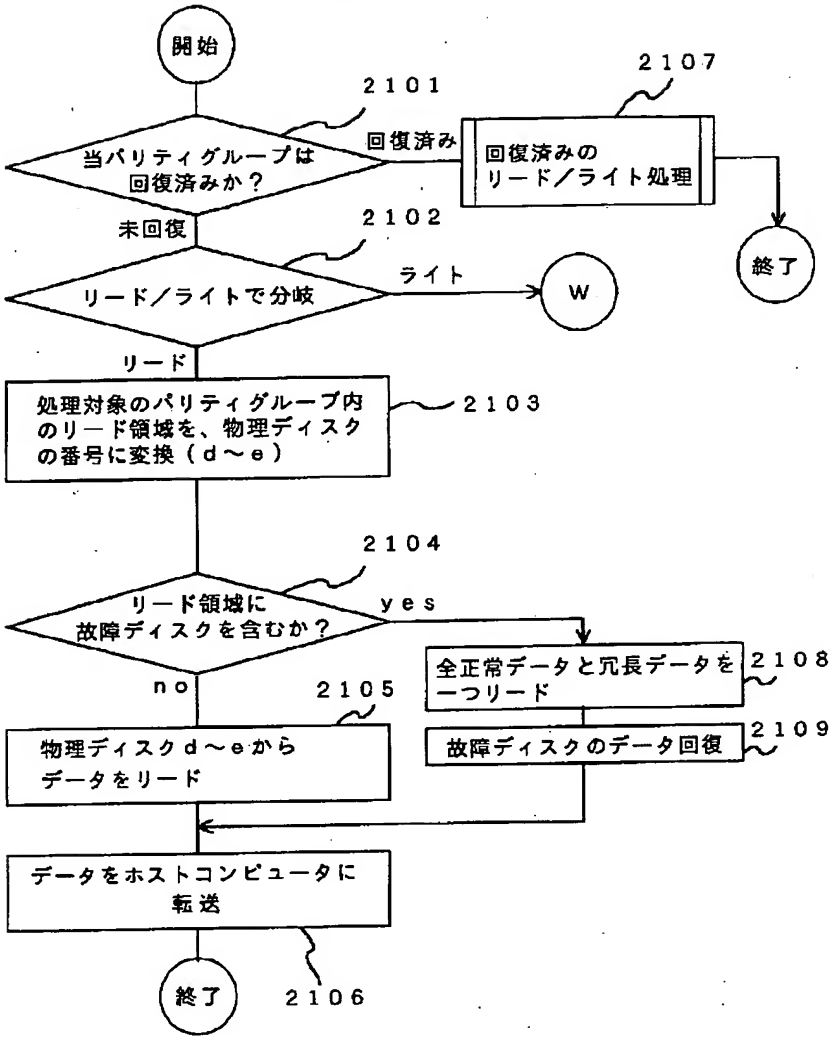
【図19】

図19


開 始
リード／ライト範囲に対応するパリティ
グループの範囲を求める。
a＝範囲内の最小のパリティグループ番号
b＝範囲内の最大のパリティグループ番号
2001
i←a
2002
i＞b？
yes
終 了
2003
no
パリティグループi内の
リード／ライト処理
2004
i＝i＋1
2005

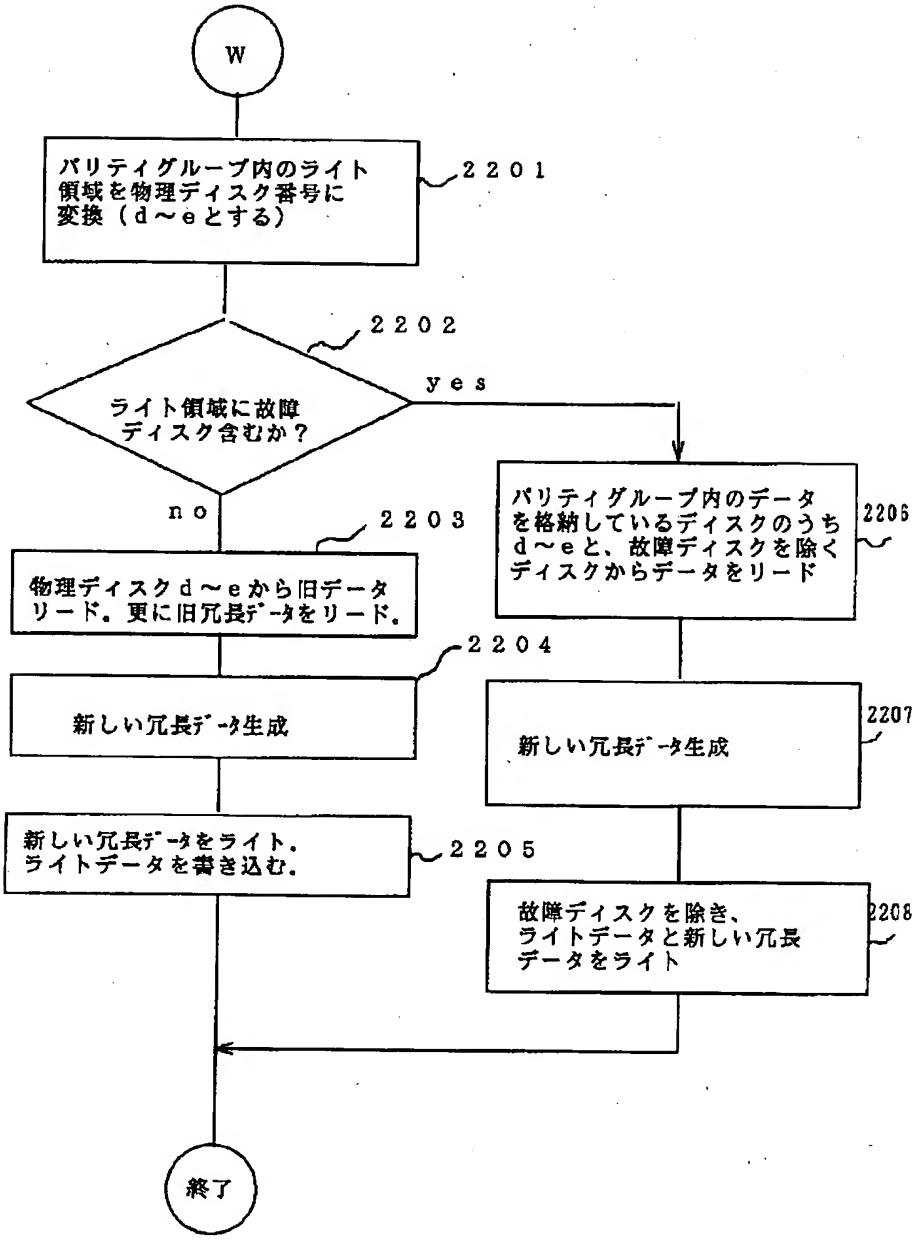
【図20】

図20


開始
2101
当パリティグループは回復済みか？
回復済み
2107
回復済みのリード／ライト処理
終了
未回復
2102
リード／ライトで分岐
ライト
W
リード
処理対象のパリティグループ内のリード領域を、物理ディスクの番号に変換（d～e）
2103
2104
リード領域に故障ディスクを含むか？
yes
全正常データと冗長データを一つリード
2108
no
2105
物理ディスクd～eからデータをリード
故障ディスクのデータ回復
2109
データをホストコンピュータに転送
終了
2106

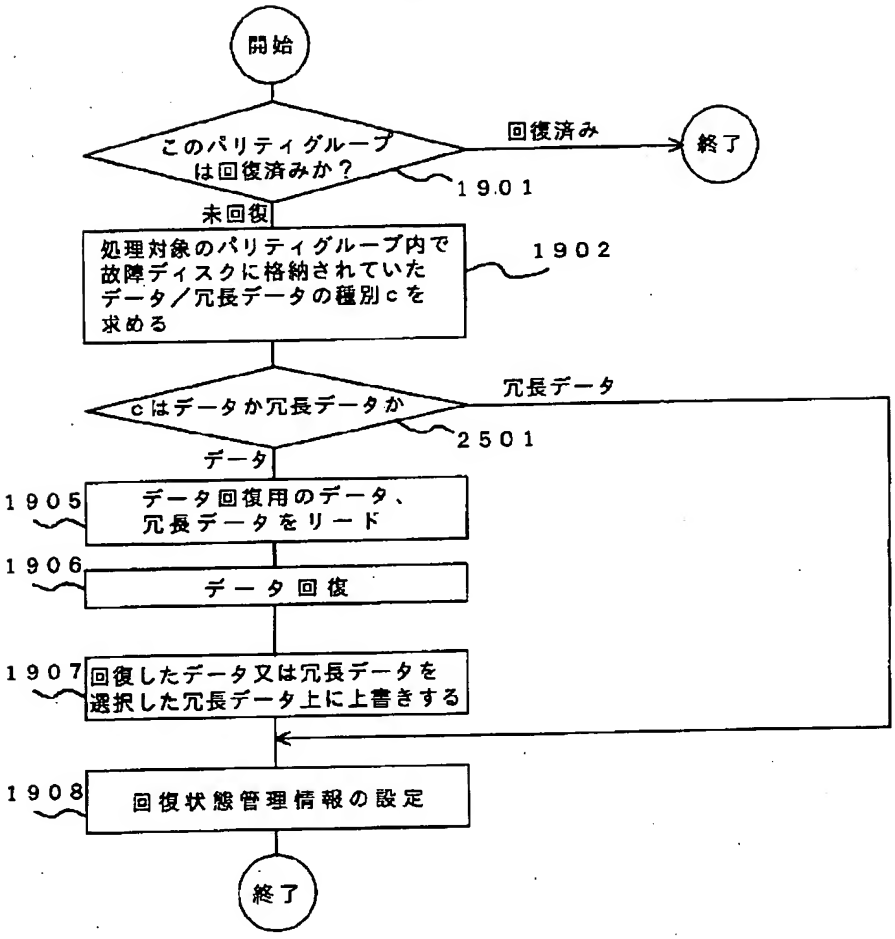
【図21】

図21


W
パリティグループ内のライト領域を物理ディスク番号に変換（d～eとする）
2201
2202
ライト領域に故障ディスク含むか？
yes
no
2203
物理ディスクd～eから旧データリード。更に旧冗長データをリード。
2204
新しい冗長データ生成
新しい冗長データをライト。ライトデータを書き込む。
2205
パリティグループ内のデータを格納しているディスクのうちd～eと、故障ディスクを除くディスクからデータをリード
2206
新しい冗長データ生成
2207
故障ディスクを除き、ライトデータと新しい冗長データをライト
2208
終了

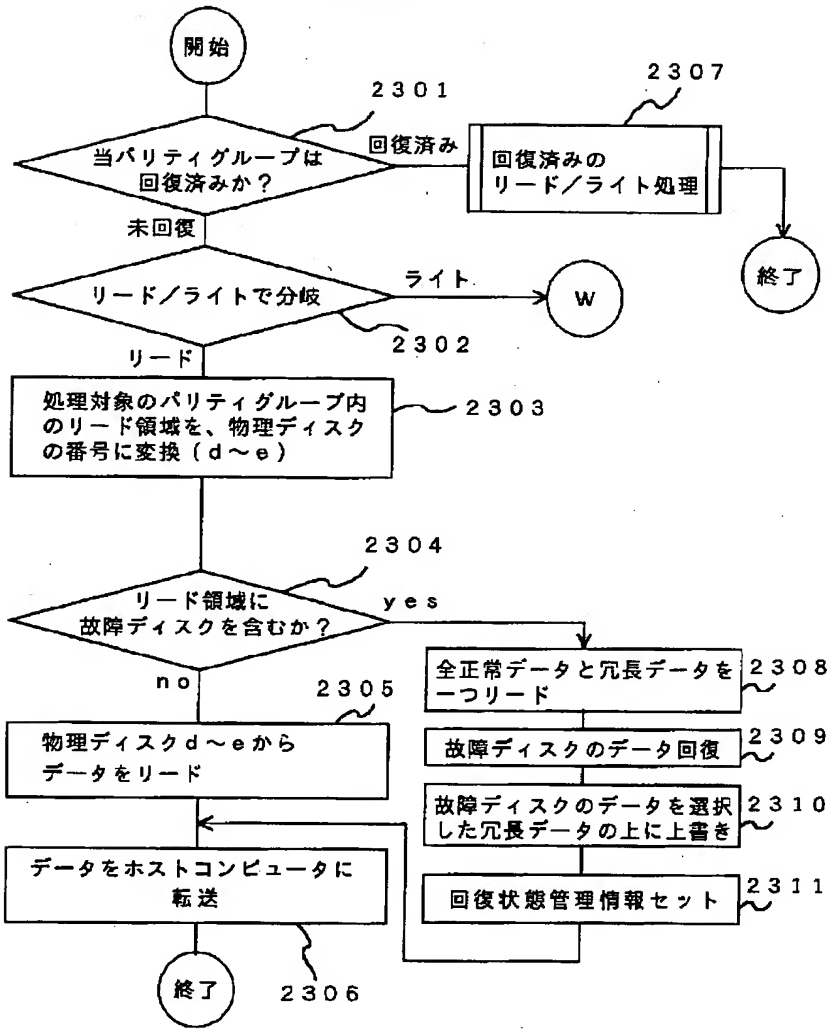
【図２２】

図２２


開始
このパリティグループは回復済みか？
回復済み
終了
１９０１
未回復
処理対象のパリティグループ内で故障ディスクに格納されていたデータ／冗長データの種別ｃを求める
１９０２
ｃはデータか冗長データか
冗長データ
２５０１
データ
１９０５
データ回復用のデータ、冗長データをリード
１９０６
データ回復
１９０７
回復したデータ又は冗長データを選択した冗長データ上に上書きする
１９０８
回復状態管理情報の設定
終了

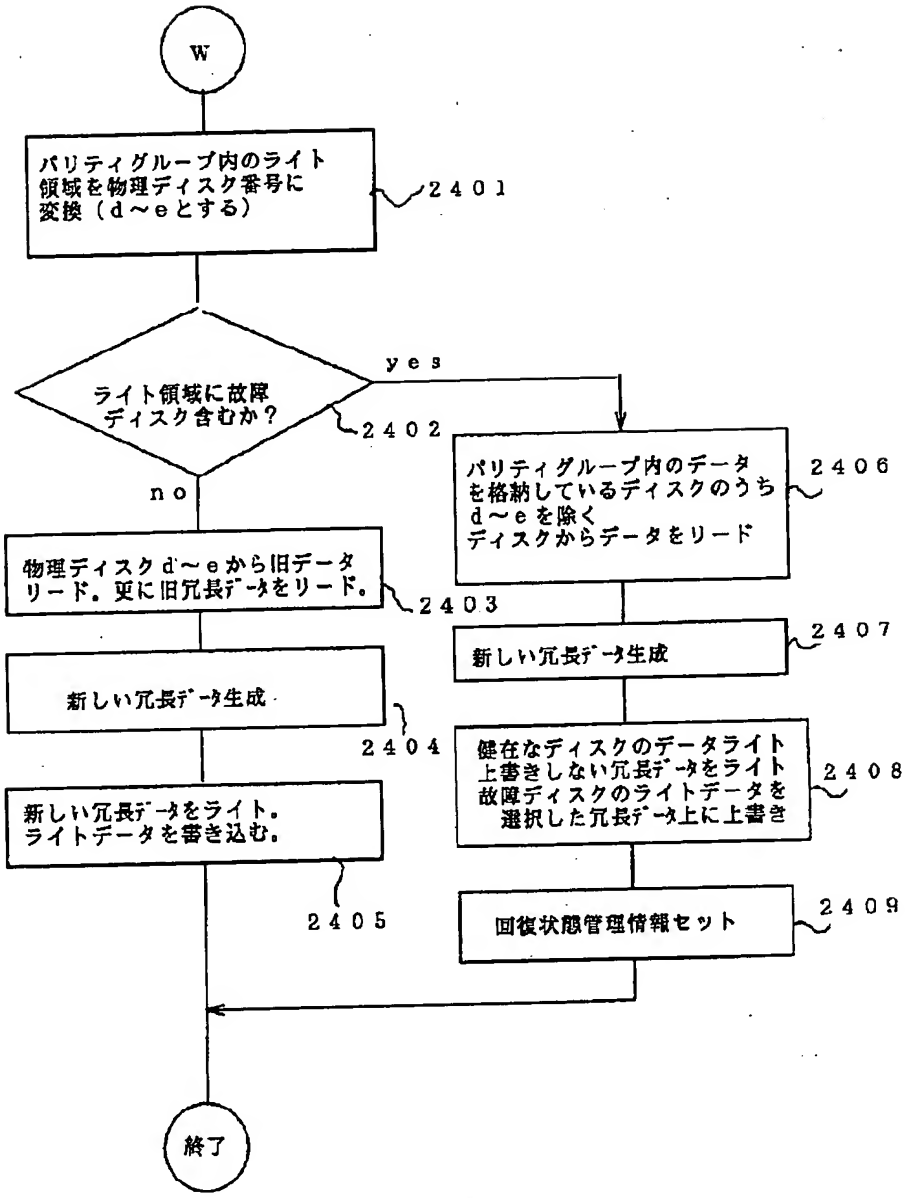
【図２３】

図２３


開始
2301
当パリティグループは回復済みか？
回復済み
2307
回復済みのリード／ライト処理
終了
未回復
リード／ライトで分岐
ライト
W
2302
リード
処理対象のパリティグループ内のリード領域を、物理ディスクの番号に変換（ｄ～ｅ）
2303
2304
リード領域に故障ディスクを含むか？
ｙｅｓ
ｎｏ
2305
物理ディスクｄ～ｅからデータをリード
全正常データと冗長データを一つリード
2308
故障ディスクのデータ回復
2309
故障ディスクのデータを選択した冗長データの上に上書き
2310
回復状態管理情報セット
2311
データをホストコンピュータに転送
終了
2306

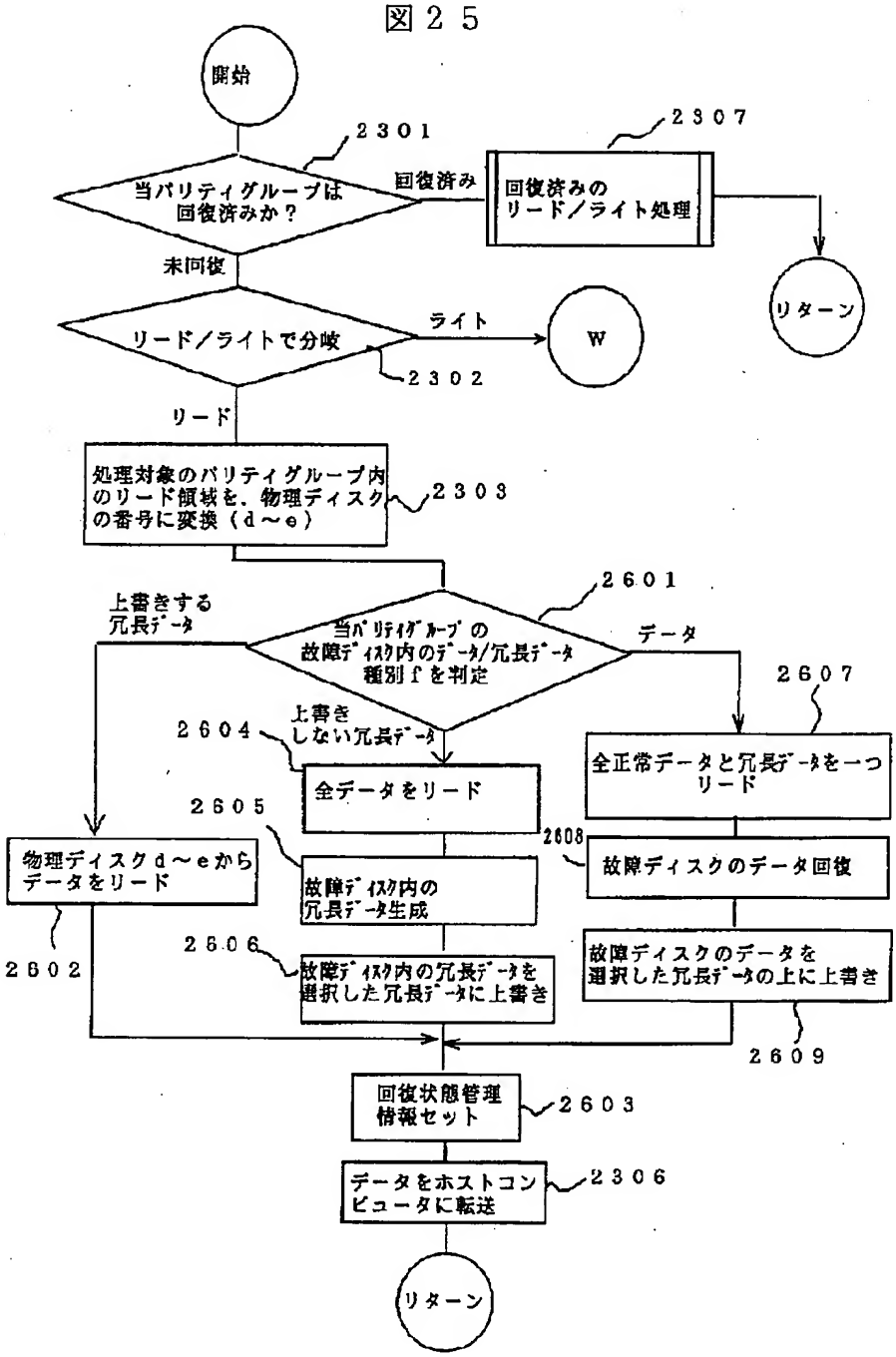
【図24】

図24


W
パリティグループ内のライト領域を物理ディスク番号に変換（d～eとする）
2401
ライト領域に故障ディスク含むか？
2402
yes
no
パリティグループ内のデータを格納しているディスクのうちd～eを除くディスクからデータをリード
2406
物理ディスクd～eから旧データリード。更に旧冗長データをリード。
2403
新しい冗長データ生成
2407
新しい冗長データ生成
2404
健在なディスクのデータライト上書きしない冗長データをライト故障ディスクのライトデータを選択した冗長データ上に上書き
2408
新しい冗長データをライト。ライトデータを書き込む。
2405
回復状態管理情報セット
2409
終了

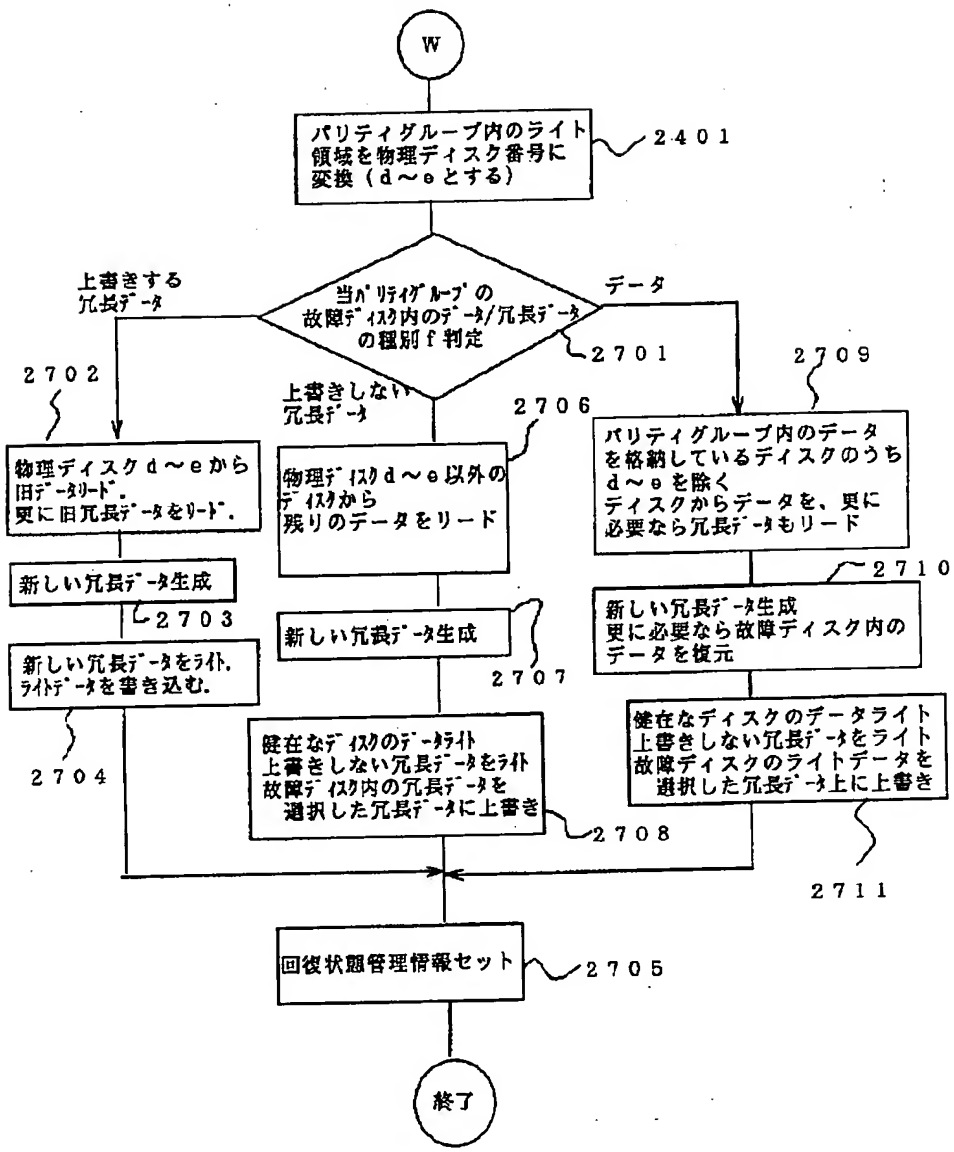
【図25】

図25


開始
2301
当パリティグループは回復済みか？
回復済み
2307
回復済みのリード／ライト処理
リターン
未回復
リード／ライトで分岐
ライト
W
2302
リード
処理対象のパリティグループ内のリード領域を、物理ディスクの番号に変換（d～e）
2303
2601
当パリティグループの故障ディスク内のデータ/冗長データ種別fを判定
上書きする冗長データ
データ
2607
2604
上書きしない冗長データ
全データをリード
全正常データと冗長データを一つリード
2605
物理ディスクd～eからデータをリード
故障ディスク内の冗長データ生成
2608
故障ディスクのデータ回復
2606
故障ディスク内の冗長データを選択した冗長データに上書き
故障ディスクのデータを選択した冗長データの上に上書き
2602
2609
回復状態管理情報セット
2603
データをホストコンピュータに転送
2306
リターン

【図26】

図26



フロントページの続き

(72)発明者　吉田　稔
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内

THIS PAGE BLANK (USPTO)